Titre: Title:	Analyse de liaison dynamique entre gènes candidats et phénotypes associés à la pression artérielle au cours de tests physiologiques
Auteur: Author:	Johanna Sandoval
Date:	2009
Type:	Mémoire ou thèse / Dissertation or Thesis
Référence: Citation:	Sandoval, J. (2009). Analyse de liaison dynamique entre gènes candidats et phénotypes associés à la pression artérielle au cours de tests physiologiques [Mémoire de maîtrise, École Polytechnique de Montréal]. PolyPublie. https://publications.polymtl.ca/130/

UNIVERSITÉ DE MONTRÉAL

ANALYSE DE LIAISON DYNAMIQUE ENTRE GÈNES CANDIDATS ET PHÉNOTYPES ASSOCIÉS À LA PRESSION ARTÉRIELLE AU COURS DE TESTS PHYSIOLOGIQUES

JOHANNA SANDOVAL
DÉPARTEMENT DE GÉNIE INFORMATIQUE ET GÉNIE LOGICIEL
ÉCOLE POLYTECHNIQUE DE MONTRÉAL

MÉMOIRE PRÉSENTÉ EN VUE DE L'OBTENTION
DU DIPLÔME DE MAÎTRISE ÈS SCIENCES APPLIQUÉES
(GÉNIE INFORMATIQUE)
JUIN, 2009

UNIVERSITÉ DE MONTRÉAL

ÉCOLE POLYTECHNIQUE DE MONTRÉAL

Ce mémoire intitulé:

ANALYSE DE LIAISON DYNAMIQUE ENTRE GÈNES CANDIDATS ET PHÉNOTYPES ASSOCIÉS À LA PRESSION ARTÉRIELLE AU COURS DE TESTS PHYSIOLOGIQUES

présenté par: SANDOVAL Johanna
en vue de l'obtention du diplôme de: Maîtrise ès sciences appliquées
a été dûment accepté par le jury d'examen constitué de:

M. PESANT Gilles, Ph.D., président
M. MERLO Ettore, Ph.D., membre et directeur de recherche
M. HAMET Pavel, M.D, Ph.D., membre et codirecteur de recherche
M. BOURDEAU Marc, Ph.D., membre

DÉDICACE

Para Francisco, por supuesto

À ma mère, mes frères et ma famille, mes sources d'inspiration, de joie, d'amour inconditionnel et de persistance, je vous dois ce que je suis.

A mes amis, sources de bon pain pour les famines d'amour, merci toujours de votre soutien.

Pendant longtemps la prière d'un non-croyant :

O. terre, cette nuit aussi tu fus constante,
Tu respires à mes pieds dans un renouveau de fraîcheur,
Déjà tu commences à m'environner de joie,
Tu éveilles et stimules en moi la ferme résolution
De tendre à jamais vers la perfection de l'existence.

LE SECOND FAUST. FRAGMENT. GOETHE

REMERCIEMENTS

Je tiens à remercier à mon partenaire de recherche, Dr. Ivan A Arenas pour toute l'aide et le support qu'il m'a fourni dès mon arrivé au Centre de Recherche du CHUM. L'ampleur de ses connaissances dans le domaine médical et génétique et son langage toujours facile à comprendre a fait que la génétique devient pour moi un centre d'intérêt plutôt qu'une question insurmontable. Merci à Pierre-Luc Brunelle sans qui ce projet ne se serait pas réalisé. Pierre-Luc a toujours été pour moi un exemple à suivre et un support inconditionnel. Merci aussi à Audrey Noël pour ses corrections, sa sollicitude et son don d'allumer la motivation des gens. Merci à mes professeurs à l'école Polytechnique qui ont semé en moi l'intérêt de tester des approches de pointe pour résoudre des problèmes complexes comme ceux qu'on connaît en bioinformatique. Merci particulièrement à Marc Bourdeau d'abord pour son sens profond de l'humanité et après pour m'avoir illuminé avec son cours de statistiques multidimensionnelles. Merci à mes collègues du CR-CHUM Mahinè Ivanga et Majid Nikpay, et à Mathieu Desnoyers de l'École Polytechnique pour tout le temps qu'ils m'ont dédié désintéressement.

Merci à mon directeur Ettore Merlo qui a cru en moi et continue de le faire. Merci Ettore de m'avoir permis d'aller vers mon rêve. Je remercie aussi au Dr. Pavel Hamet pour m'avoir donné la chance d'appartenir à son équipe réputée de chercheurs, c'est un honneur et un défi immense et j'espère être toujours à la hauteur de ses attentes.

Finalement, je remercie à Génome Canada pour son support dans la présentation de mon affiche à HUGO 2007

RÉSUMÉ

La pression artérielle (PA) est un caractère complexe qui semble être lié à plusieurs interactions gène/environnement. La mise en évidence de la liaison entre certaines régions chromosomiques et les traits intervenant dans la variation de la PA sur des individus soumis à des stimuli physiologiques devrait permettre d'élucider d'autres mécanismes responsables de certains désordres vasculaires et leurs complications. Primo, nous postulons que le degré de liaison génétique varie selon les tests physiologiques. Ensuite, nous postulons que la réduction de dimensionnalité des phénotypes à l'aide de méthodes reflétant la variance et la corrélation entre les traits permettra d'améliorer les signaux des tests d'association. 258 individus sélectionnés d'une cohorte de familles du Saguenay-Lac-Saint-Jean au Québec ont été soumis à certaines manœuvres orthostatiques et des traits reliés à la pression artérielle ont été mesurés en utilisant l'impédance cardiaque. Les pressions systolique (SBP) et diastolique (DBP), la pression artérielle moyenne (MAP), la résistance vasculaire totale (TPR), le pouls (HR) et le volume systolique (SV) ont été mesurés 6 fois chaque 5 minutes pendant que les individus adoptaient la position couchée et 6 fois chaque 2 minutes pendant la position debout. 1000 marqueurs génétiques ont été soumis à des tests de liaison génétique à chaque temps de mesure. Ces marqueurs sont sur 361 gènes révélés dans des travaux antérieurs comme étant engagés dans la fonction cardiaque et vasculaire. Pour tester la première hypothèse et pour repérer les marqueurs génétiques qui présentent ce comportement, nous avons implanté un test de permutation sur la corrélation entre la condition expérimentale et les signaux de liaison obtenus pour les marqueurs et sur la différence des moyennes des signaux de liaison dans les deux périodes. Ensuite, nous avons utilisé une méthode de rééchantillonnage pour ajuster la signification des valeurs p des analyses de liaison par période ainsi que pour quantifier la liaison dynamique par rapport aux phénotypes, les génotypes et les familles testées. Pour tester la deuxième hypothèse, nous avons comparé les méthodes de la moyenne, l'analyse factorielle (AF) et l'analyse des composantes principales (ACP) dans le but de produire des phénotypes unidimensionnels qui ont été soumis aux tests d'association avec FBAT (*Family based association tests*). Les valeurs p issues des tests d'association ont été comparées avec des tests des différences des moyennes. Pour les deux périodes FA, PC et la méthode de la moyenne ont fourni des meilleurs résultats que la méthode d'ajustement traditionnelle (Bonferroni). Nous avons implanté une simulation nous permettant de confirmer que les méthodes choisies contrôlent l'erreur de type I attendu. Cette simulation-ci nous a permis aussi de relaxer les seuils de signification en calculant une valeur approximative du seuil de signification selon les données observées. Les résultats observés par simulation nous ont

permis de confirmer que les techniques de réduction de dimensionnalité reflétant la corrélation et la variance commune entre les phénotypes intermédiaires de la pression artérielle nous ont permis d'améliorer les résultats des tests d'association. Les tests de liaison durant des tests physiologiques (liaison dynamique) ont rapporté quelques marqueurs et régions chromosomiques dernièrement répertoriés pour être associés à des processus cardiovasculaires. Les marqueurs génétiques liés dynamiquement non répertoriés pourraient être reliés aux facteurs génétiques qui seront à l'origine de réponses physiologiques.

ABSTRACT

Blood pressure (BP) is a complex trait resulting from several gene-environment interactions. It is likely that loci (a fixed position on a chromosome) involved in BP variation under physiological stimuli may also be involved in the development of blood pressure disorders and its complications. First, we hypothesized that the degree of genetic linkage varies with physiological responses. Second, we hypothesized that reduction techniques which reflect the common variance or the correlation between blood pressure intermediate phenotypes may improve the results of association tests. 258 individuals selected from a cohort of French Canadian families were subjected to orthostatic maneuvers, and systolic (SBP), diastolic (DBP) and mean arterial BP (MAP) and intermediate phenotypes (total peripheral resistance (TPR), heart rate (HR) and stroke volume (SV) were measured every 5 and 2 minutes during supine (30min) and after adopting the standing position (10 min) using cardiac impedance. Standing was associated with drastic changes in all phenotypic means and an increased variance compared with the supine period. BP was highly correlated with TPR and moderately correlated with SV and HR, whereas TPR and SV were also highly correlated. 1000 candidate genetic markers related to cardiac and vascular function were tested for genetic linkage at each measure in time. To test the first hypothesis, and to find the genetic markers that give rise to the effect of physiological stimuli, we used a step-down permutation test on the correlation between the experimental condition and the linkage statistics and on the difference between the linkage statistics into the different periods. We used a resampling based strategy to assess the significance of the linkage in each period and to confirm the dynamic linkage in some interesting markers. To test the second hypothesis we compared the average method, principal components (PC), and factor analysis (FA) to generate univariate traits for FBAT association studies during physiological testing on more than 300 markers selected from the linkage tests. The overall P-values obtained from FBAT with each method were compared using a means difference test. For both the supine and standing periods FA, PC and average method provided better overall FBAT p-values compared with the traditional Bonferroni adjustment. Using a permutation procedure we confirmed that the selected methods controls the type-I error and we compute an approximate signification threshold for each method. These observations suggested that reduction techniques which reflect the common variance or the correlation between blood pressure intermediate phenotypes may improve the results of association tests. Linkage analysis during physiological testing (dynamic linkage) reported significant markers and complete chromosomal regions lately reported as related to cardiovascular processes. On unknown or newly discovered regions dynamic linkage may uncover genetic factors that drive physiological responses.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES SIGLES ET ABRÉVIATIONS

ACP	Analyse des composantes principales
AF	Analyse factorielle
CR-CHUM	Centre de Recherche du Centre Hospitalier de l'Université de Montréal
DIA, DBP	pression artérielle diastolique (de l'anglais *diastolic blood pressure*)
FA	Analyse factorielle (de l'anglais *factor analysis*)
HT	hypertension
HE	Modèle de Haseman-Elston pour l'analyse de liaison génétique
HWE	équilibre d'Hardy-Weinberg
LD	déséquilibre de liaison
MAP	pression artérielle moyenne (de l'anglais *mean arterial blood pressure*)
MmHg	millimètres de mercure
PAD	pression artérielle diastolique
PAS	pression artérielle systolique
bp	paires de bases
PTM	Problème de tests multiples
QTL	quantitative trait loci
SBP, SYS	pression artérielle systolique (de l'anglais *systolic blood pressure*)
SLSJ	Saguenay-Lac-Saint-Jean
SNP	polymorphisme d'un nucléotide simple (de l'anglais *single nucleotide polymorphism*)
SV	volume d'expulsion (de l'anglais *stroke volume*)
TDT	test du déséquilibre de transmission
TPR	résistance vasculaire totale (de l'anglais *total peripheral resistance*)
PCA	Analyse des composantes principales (de l'anglais *Principal component analysis*)

LISTE DES ANNEXES

CHAPITRE 1. INTRODUCTION

Il est couramment accepté que les maladies complexes sont causées par des facteurs environnementaux combinés à des facteurs génétiques amenés non pas par un seul gène agissant seul, mais par plusieurs gènes qui interagissent les uns avec les autres. En raison du grand nombre de marqueurs génétiques disponibles dans le génome entier, la charge du calcul pour les effets marginaux et pour toutes les paires, triplets, et même des interactions d'ordre supérieur est pour l'instant détournée par des approches simples et par des tests unidimensionnels. Une de ces approches consiste à identifier un plus petit nombre de SNP positionnés sur des gens candidats, en utilisant l'analyse de liaison. Avec une liste de marqueurs génétiques raffinée, une analyse statistique plus approfondie peut être effectuée. Lors de cette deuxième étape, un test unidimensionnel d'association en présence de liaison est couramment utilisé.

Pour déterminer une liaison génétique, un test statistique examine directement la transmission intergénérationnelle des deux phénomènes: le phénotype, résultant dans une maladie, et les allèles d'une famille, en cherchant des corrélations qui suggèrent que le marqueur génétique est lié à un locus (un emplacement sur un chromosome) causal. Si les individus qui ont hérité la maladie héritent aussi l'information dans le marqueur génétique, alors le gène qui cause la maladie et le marqueur ont tendance à être proches dans le chromosome.

Dans une étude de liaison ou d'association on peut tester plusieurs phénotypes, utiliser plusieurs types de tests et explorer plusieurs modèles génétiques. Ainsi, on effectue plusieurs tests statistiques et on est devant un problème de tests multiples. La raison qui justifie la correction par tests multiples est qu'en appliquant plusieurs tests d'hypothèse à la fois on augmente la probabilité de déclarer des résultats faussement significatifs, c'est-à-dire, associations statistiquement significatives qui ne le sont pas. Par exemple, dans une étude particulière les chercheurs testent quelques hypothèses et trouvent un résultat significatif pour l'une d'elles avec une valeur p = 0.005. Cette valeur p est interprétée comme suit: lorsque la relation de causalité n'existe pas pour l'effet testé (l'hypothèse nulle est vraie), il y a une probabilité de 0.05% d'observer un résultat aussi extrême que le résultat observé. Mais devant le test simultané de plusieurs hypothèses nulles on a le risque que 5% des tests rapporteront une valeur p inférieur ou égal à 0.05 et la plupart de ces associations déclarées comme

significatives le sont faussement. Un ajustement de la valeur p par test multiples permet de contrôler le taux de faux positifs et la nouvelle valeur p peut être interprétée comme suit: lorsque la relation de causalité n'existe pas pour n'importe lequel des effets testés, il y a une probabilité de p_{adj} (étant p_{adj} la valeur p ajustée du test) que quelque part dans l'expérience on observe un résultat aussi extrême que le résultat observé de p.

L'ajustement de Bonferroni est une procédure générique de résolution du problème des tests multiples dont le seuil de signification devient $\alpha_{adj} = \alpha/m$, ce qui équivaut à ajuster chaque valeur p en le multipliant par m, le nombre de tests effectués. Donc, une hypothèse nulle sera rejetée par la méthode de Bonferroni si sa valeur p ajustée p_{adj} est inférieur à α. La correction de Bonferroni s'avère conservatrice lorsque les tests ne sont pas indépendants. Et dans le cas particulier des études génétiques des corrélations sont présentes par plusieurs raisons. D'abord, les marqueurs qui sont rapprochés peuvent être corrélés (effet connu comme déséquilibre de liaison). Ensuite les mesures répétées des phénotypes sont corrélés et les phénotypes définissant une maladie complexe peuvent aussi être corrélés. D'autres méthodes qui tiennent compte de la corrélation des tests on été développées, mais aucune méthode n'est plus prometteuse que les méthodes de rééchantillonnage. Les méthodes de rééchantillonnage permettent d'utiliser les données observées exhaustivement dans le but de faire des inférences. Les variables observées sont donc réassignées aléatoirement aux individus en étude et les statistiques sur ces nouveaux arrangements sont recalculées des milliers de fois. La valeur de la statistique observée est donc considérée inusuelle si elle est inusuelle par rapport à la distribution calculée par rééchantillonnage. Des méthodes basées sur le rééchantillonnage ont été développés pour les études familiales, mais aucune étude ne traite les mesures répétées des phénotypes.

La population cible dans cette étude est celle du Saguenay-Lac-St-Jean (SLSJ), reconnue comme une population à effet fondateur. L'équipe de recherche du Dr. Pavel Hamet s'intéresse à élucider la composante génétique de l'hypertension. La pression sanguine, trait complexe qui caractérise l'hypertension, change continuellement en réponse à divers stimules environnementaux autant internes que externes, tels que la position de la personne, la présence ou absence d'un examinateur, l'heure du jour, etcétéra. Les effets environnementaux augmentent la variabilité de la pression sanguine et diminuent les chances de détecter des vraies liaisons ou associations génétiques. D'autre part en état de repos plusieurs mécanismes physiologiques interactifs qui maintiennent la pression sanguine peuvent se chevaucher et pourraient cacher des petits effets génétiques qui seraient seulement identifiables après un

certain stimule. Mais avant tout on croit que les facteurs qui contrôlent les réponses de la pression sanguine sont aussi impliqués dans pathogénèse de l'hypertension.

Le problème de liaison dynamique ne s'est jamais posé comme tel dans la littérature avant nous. Un problème qui lui ressemble a été proposé dans l'atelier de génétique humaine GAW13 en 2003 dans le cadre d'une étude sur les phénotypes longitudinaux, c'est-à-dire, ceux dont la valeur est mesurée pendant des intervalles de la vie des individus. L'atelier avait comme base des données de l'étude Framingham Heart, une initiative très reconnue dans le domaine cardiovasculaire crée dans le but d'élucider les risques reliés aux maladies cardiovasculaires. Des solutions bases sur la réduction unidimensionnelle ont été proposées, et les créateurs de FBAT ont répondu en produisant un test (FBAT-PC) qui applique la réduction de dimensionnalité par une combinaison de l'analyse de composantes principales et l'héritabilité des phénotypes (Lange, et al., 2004). Malheureusement, la structure de nos familles est tellement complexe et dans quelques cas les familles sont si grandes que le programme n'est pas capable de réaliser les tests. Nous avons donc réalisé nous-mêmes les réductions de dimensionnalité sur les phénotypes et par la suite nous avons analysé les effets des réductions par simulations.

Nous proposons plusieurs analyses qui utilisent les méthodes de rééchantillonnage. D'abord nous utilisons un test de permutation pour confirmer que les statistiques du test de liaison varient selon les stimuli environnementaux externes. Ensuite nous utilisons une méthode de permutation pour quantifier la signification de l'effet des tests physiologiques par rapport aux marqueurs disponibles, la généalogie et les phénotypes mesurés. Finalement nous proposons plusieurs méthodes de réduction de dimensionnalité des phénotypes pour tester l'association entre les phénotypes intermédiaires de l'hypertension et l'ensemble des marqueurs significativement liés détectés dans la premier partie de l'étude. Des simulations sont effectués pour vérifier le control de l'erreur de type I dans les tests d'association et pour ajuster seuil de signification par rapport aux données observées.

1.1 Définition du problème

En exploitant la disponibilité d'une source importante d'information de phénotypes dynamiques liés à la variation de la pression artérielle dans une population génétiquement homogène, l'objectif du présent travail est d'identifier et de quantifier l'effet des tests physiologiques sur la

liaison génétique. Cette quantification devrait apporter à l'équipe de génomique prédictive du Centre de Recherche du CHUM des pistes sur les processus complexes qui déclenchent l'hypertension. Il s'agit donc de rechercher et l'implanter des modèles statistiques convenables aux données et d'appliquer des tests basés sur ces modèles à un cas d'étude particulier sur lequel nous avons été assignés: de quantifier la liaison entre certains marqueurs génétiques et des multiples traits corrélés, de confirmer et de mesurer la composante dynamique de cette liaison par rapport aux tests physiologiques. Par la suite, ayant établi l'effet des tests physiologiques sur la liaison génétique nous voulons aussi utiliser des méthodes de réduction de dimensionnalité des variables phénotypiques telles que l'analyse de composantes principales, l'analyse factorielle et les statistiques descriptives dans le but de diminuer l'effet des tests multiples sur la signification des tests d'association génétique tout en contrôlant le nombre de fausses associations rapportées .

1.2 Division du mémoire

Ce mémoire se divise ainsi. Au chapitre 2 nous présentons quelques concepts génétiques qui servent à comprendre les tests de liaison et association génétique. Dans le chapitre 3 nous faisons un rappel sommaire de certaines notions statistiques, étant la statistique une composante très importante de la solution des problèmes présentés dans ce mémoire. Ensuite, dans le chapitre 4 nous présentons les tests de liaison et d'association génétique. Nous décrivons dans le chapitre 5 le problème des tests multiples, qui nous ont conduits au choix des méthodes de rééchantillonnage et réduction de dimensionnalité dans l'étude ici présenté. Dans le chapitre 6 nous faisons un rappel du problème, à la manière des textes scientifiques en génétique et nous présentons les méthodes proposés en détaille. Nous présentons au chapitre 7 les critères qui ont guidé le développement du logiciel, les activités de génie logiciel suivies et les méthodes proposées à un plus haut niveau de détaille. Au chapitre 8 nous décrivons les expériences que nous avons achevées et les résultats que nous avons obtenus, pour en finir au chapitre 9 avec une discussion, les conclusions et des indices proposés pour des travaux futurs.

CHAPITRE 2. CONCEPTS GÉNÉTIQUES

Dans ce chapitre nous faisons un rappel des notions génétiques courantes dans le langage des tests d'association et de liaison. Nous touchons aux concepts suivants: la démarche scientifique qui a amené à la découverte des gènes, les chromosomes, l'ADN, les marqueurs génétiques, les mesures de la diversité génétique des populations, une brève introduction à la définition de maladie complexe et les ressources dont ont se sert pour valider la viabilité biologique des résultats obtenus dans les tests de liaison et d'association génétique.

2.1 Les gènes

La génétique telle qu'on la connaît actuellement est l'étude des gènes à tous les niveaux, de la molécule aux populations. Cette discipline a été initiée avec le travail de Mendel, qui a formulé l'idée que les gènes existent, sans pourtant les nommer comme tels. Nous savons maintenant qu'un gène est une région fonctionnelle d'une longue molécule (l'ADN) qui constitue la structure fondamentale des chromosomes. Plusieurs gènes codent la structure d'une protéine et les protéines déterminent les propriétés d'un organisme. Chaque produit d'un gène participe donc à la réalisation de multiples caractères et chaque caractère résulte de l'action de centaines, voire de milliers de protéines(Schalchli, 2005).

Depuis toujours, nous présumons que certains caractères physiques sont héréditaires vu la ressemblance entre les parents et les enfants. Pourtant, ce n'est qu'à la fin du XXe siècle que le mécanisme de l'hérédité a été révélé. Jusqu'au XIXe siècle c'était le modèle de la théorie hippocratique de la génération qui prévalait, et qui est reprise par Darwin sous le nom de " pangenèse ". Chez Hippocrate, le corps est constitué de différentes humeurs dont un échantillon est envoyé pour chaque partie du corps aux organes génitaux pour former les semences. Ces semences se mélangent après la fécondation et produisent, à partir des humeurs des parents, la tête, les bras, etc. L'enfant est ainsi construit à partir d'un échantillon représentatif des humeurs de ses parents et présente donc les mêmes caractères qu'eux. Auguste Weismann, dans sa théorie du plasma germinatif, a fait remarquer que le spermatozoïde et l'ovule, cellules reproductrices, sont différenciés très précocement au cours du développement embryonnaire. Les autres cellules du corps, appelés somatiques, bien qu'elles

dérivent aussi des cellules germinales ne peuvent pas les engendrer à nouveau. Il a réfuté alors les idées de pangenèse (Pichot, 2005).

Au cours des années 1860, les noyaux des cellules eucaryotes[1] ont pu être observés sous forme de corpuscules allongés appelés chromosomes. Normalement, il y a deux copies de chaque chromosome sous forme de paires homologues dans chaque cellule somatique. Contrairement aux cellules somatiques, chacune des cellules reproductrices, ou gamètes, n'a que 23 chromosomes : un jeu de 22 autosomes et un chromosome sexuel, soit X, soit Y. On appelle cellule haploïde une cellule qui n'a qu'un seul jeu de chromosomes. Chez l'humain, le nombre haploïde (N) est 23. L'œuf résultant de la fécondation (zygote) et toutes les autres cellules qui possèdent deux jeux de chromosomes sont des cellules diploïdes. Chez l'humain, le nombre diploïde (2N) est 46.

En 1866, Gregor Mendel publie les lois fondamentales de l'hérédité obtenues à partir des résultats de ses expériences d'hybridation de petits pois dans le jardin de son monastère. Les observations de Mendel sont interprétées en faisant l'hypothèse que les différentes paires de caractères alternatifs résultent chacune de l'action d'un facteur (appelé plus tard gène) qui possède des formes alternatives (allèles). Chaque plante posséderait donc une paire de gènes déterminant un caractère particulier, dont un exemplaire est hérité de chacun de ses parents (Voet & Voet, 1998). Les lignées distinctes (ou individus) représentent alors des formes différentes qui peuvent prendre les caractères : on les nomme formes de caractères, variantes pour un caractère ou **phénotypes**. La première loi, la loi de ségrégation, est découverte par le croisement de deux lignées de pois différant par un caractère seulement. La première génération comprend des pois hybrides et d'aspect lisse. En permettant l'autofécondation, la deuxième génération donne les proportions suivantes : ¼ de pois ridés et ¾ de pois lisses. Les mêmes proportions sont observées à chaque autofécondation. La deuxième loi, l'assortiment indépendant des caractères, est observée en croisant deux couples de caractères, soit l'aspect et la couleur, et en observant la proportion relative des quatre types de pois (jaune ou vert, lisse ou ridé). Des proportions de 9:3:3:1 amènent à Mendel à déduire que la transmission des caractères est indépendante. Si cela n'avait pas été le cas, les croisements n'auraient donné

[1]Il existe deux types fondamentaux de cellules selon qu'elles possèdent ou non un noyau: les eucaryotes et les procaryotes

que deux types de descendance (étant donné la dominance des traits) dans les proportions ¾, ¼.

Le contexte scientifique du début du XX siècle va changer radicalement l'approche de l'hérédité. En 1900 on redécouvre les lois de Mendel, qui n'étaient jusqu'à là que des lois d'hybridation. En 1902, Sutton remarque que le comportement des caractères mendéliens est analogue au comportement des chromosomes lors de la mitose et de la méiose[2], processus de division cellulaire dont l'utilité dans le cycle de vie d'un organisme diploïde est montrée par la Figure 2.1.

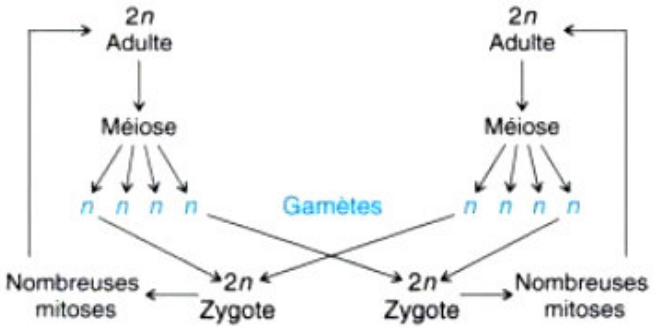


Figure 2.1. Le cycle de vie d'un organisme diploïde, tel que l'humain tiré de (Griffiths & Suzuki, 2002).

Sutton formule alors la théorie chromosomique de l'hérédité, dans laquelle il propose que les gènes sont des parties de chromosomes. Viennent ensuite les travaux de Thomas Morgan sur les mouches du vinaigre (*Drosophile*). Morgan propose un modèle qui localise les gènes (plus exactement les mutations) sur les différents chromosomes. Le gène est alors défini comme un "**locus**", un emplacement sur un chromosome. Après les mouches, les études sur les microorganismes se révèlent fécondes. En 1941, Beadle et Tatum découvrent que chaque gène semble gouverner la synthèse d'une enzyme (protéine), sans pourtant savoir à quoi ressemblent les fameux gènes. En 1953, Crick et Watson élucident la structure de l'ADN. Une double hélice formée de deux brins portant des bases complémentaires, ce qui explique comment il peut être copié à l'identique et transmis aux cellules fille.

[2] La mitose assure la naissance de cellules identiques à la cellule mère lors de la reproduction asexuée. La méiose produit les cellules sexuelles ou gamètes pour la reproduction.

Nous savons maintenant qu'un gène est une région fonctionnelle d'une longue molécule (l'ADN) qui constitue la structure fondamentale des chromosomes. L'ADN est composé de plusieurs substances chimiques, contenant toutes du sucre, un groupement phosphate et une des quatre bases – l'adénine (A), la thymine (T), la guanine (G), et la cytosine (C). L'ADN est donc formé de deux brins portant des bases complémentaires liées (A avec T et G avec C). Lors de la division cellulaire, les deux chaînes se séparent pour permettre la synthèse de deux molécules filles identiques d'ADN.

Plusieurs gènes codifient la structure d'une protéine et les protéines à leur tour déterminent les propriétés d'un organisme. Du gène à la protéine, des nombreuses étapes se suivent. L'ADN est transcrit en ARN pré messager qui va subir diverses maturations. Celle qui détermine la production des protéines s'appelle l'épissage. Dans l'épissage, des segments de cet ARN (les introns) sont excisés et ceux qui subsistent (les exons) sont mis bout à bout. Ils constituent une nouvelle molécule plus courte, l'ARN messager (ARNm). Cet ARNm sort du noyau et sert de trame lors de la formation des protéines. Des 1957 Crick énonce ce qui allait devenir le **dogme central** de la biologie moléculaire : « l'ADN fabrique l'ARN, L'ARN fabrique des protéines et les protéines nous fabriquent " (Schalchli, 2005). Nous savons aujourd'hui que les processus de transcription et de traduction sont beaucoup plus complexes : la découverte des gènes régulateurs, dont le produit est une protéine capable de réguler l'activité d'autres gènes et le morcellement de certains gènes dans le génome fait que dans certaines circonstances un gène donne plusieurs protéines. Tout dépend du type de cellule ou encore de son état (stades du développement). Chaque produit de gène participe à la réalisation de multiples caractères et chaque caractère résulte de l'action de centaines, voire de milliers de protéines.

Le concept de gène est alors devenu pluriel: un fragment d'ADN transmis de génération en génération fonctionnellement affecté par la cellule en fonction du lieu et du moment, qui va aboutir à la synthèse d'une protéine donnée. Chaque gène occupe un emplacement précis dans un chromosome, tel un livre dans une bibliothèque. Chez l'humain, les 46 chromosomes vont par paires chacune étant héritée du père et de la mère. La paire numéro 23 est différente puisqu'elle présente deux chromosomes distincts: un chromosome X et un Y chez les garçons et une paire de chromosomes X chez les filles. Chaque gène est donc présent en double exemplaire sous forme d'allèles (sauf pour les gènes présents dans les chromosomes X et Y chez les garçons). Un individu est dit homozygote pour un gène lorsque les deux allèles du gène en question sont identiques, et hétérozygotes s'ils ne le sont pas. L'expression coordonnée des gènes dans le temps et dans l'espace permet la synthèse de différentes protéines qui, en

formant des voies métaboliques, seront à l'origine des processus physiologiques tels que le rythme cardiaque, la formation des organes, etc. (Schalchli, 2005)

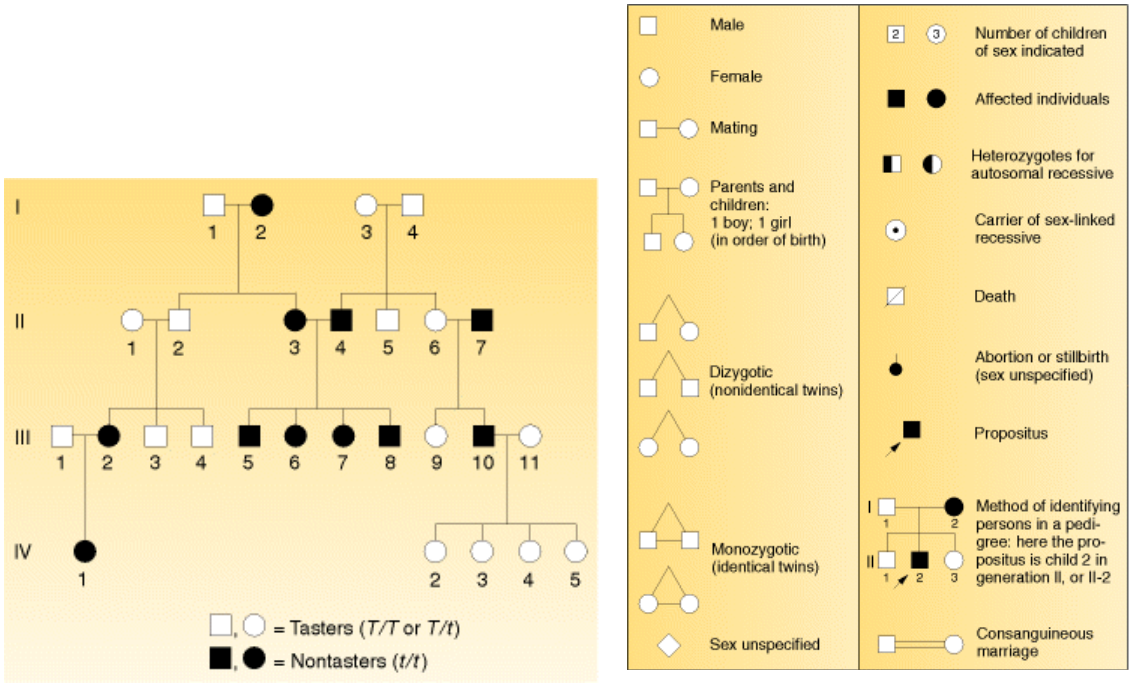
2.2 La variation génétique

La variation génétique pour les membres d'une espèce, étant donné qu'ils possèdent les mêmes gènes, est donnée par l'existence de différentes formes d'un gène, c'est-à-dire, les allèles. La constitution allélique d'un organisme est son génotype. Cependant, on utilise aussi le mot pour parler d'un seul gène. Puisque tous les êtres humains possèdent deux jeux de chromosomes dans chaque cellule, les génotypes peuvent être A/A, A/a ou a/a. Les allèles *A* et *a* sont les allèles " type " d'un même gène. Les cellules haploïdes peuvent être de génotype *A* ou *a* et les diploïdes peuvent être homozygotes, *A/A* ou *a/a*, ou hétérozygotes, *A/a*. Les allèles d'un gène sont tous dans la même position chromosomique. Un allèle peut être soit dominant, soit récessif. La dominance est définie lorsque la présence d'un allèle sur seulement l'un des deux chromosomes fait qu'un caractère soit exprimée. Un allèle est récessif si son expression nécessite sa présence sur les deux chromosomes (Griffiths & Suzuki, 2002)

La variation allélique est à la base de la variation héréditaire. Une classification des types de variation très répandue divise la variation en deux classes: continue et discontinue. Dans la variation **discontinue,** un caractère donné existe sous formes distinctes ou phénotypes. On cherche à démontrer que les phénotypes alternatifs sont codés par les allèles d'un même gène. Par exemple dans l'albinisme le phénotype de *A/A* est pigmenté, *a/a* est albinos et *A/a* est pigmenté. La présence de différences alléliques entraîne des différences phénotypiques, ce qui ne signifie pas qu'un seul gène affecte la couleur de la peau. La variation **continue** d'un caractère (tel que la taille, le poids ou l'intensité de la couleur) présente une gamme ininterrompue de phénotypes dans la population. Lorsque les fréquences phénotypiques sont représentées sous la forme d'un graphe, on observe une distribution en cloche. Le plus souvent, la variation est à la fois d'origine génétique et environnementale et dans certains cas, soit génétique, soit environnementale. On cherche donc à trouver la proportion de chaque composante (Griffiths & Suzuki, 2002).

Un grand nombre de maladies sont déterminées par des allèles récessifs ou dominants de gènes situés sur les chromosomes autosomiques (différents des chromosomes sexuels X et Y). Ces allèles sont transmis strictement suivant un mode mendélien. Les modes de transmission

héréditaire peuvent tous se déduire de l'analyse d'arbres généalogiques, en suivant certaines lois standards. La Figure 2.2 (b) présente les règles standard d'interpréter un arbre. Les individus atteints sont représentés par un symbole coloré. Le symbole associé aux individus de sexe masculin est le carré et celui du sexe féminin est le cercle. La descendance des couples est imagée par un trait vertical. La Figure 2.2 (a), montre la transmission de la capacité de percevoir le goût d'une substance chimique, le phénylthiocarbamide (PTC). Le phénotype devient la capacité ou incapacité pour détecter le goût et le génotype est donné par la bande inférieure. D'après l'arbre, deux goûteurs peuvent parfois engendrer des enfants non goûteurs. L'allèle de détection du goût est dominant et l'allèle de l'incapacité correspondante est récessif. Un autre exemple, la fibrose kystique du pancréas dans la population du Québec, est associé dans la majorité des cas à l'apparition de plusieurs allèles mutés du gène CTFR (*cystic fibrosis transmembrane conductance regulator*), localisé sur le chromosome 7.



(a) **(b)**

Figure 2.2 Arbre généalogique de la capacité à percevoir le goût du PTC (a) et symboles utilisés dans l'analyse d'arbres généalogiques humains (b) (Griffiths & Suzuki, 2002)

2.3 Les marqueurs

Le polymorphisme génétique est un concept très important en génétique médicale. Une variante (allèle) d'un gène est désignée comme **polymorphique** s'il existe à une fréquence de plus de 1% dans une population. Une mutation par contre est une séquence d'ADN qui diffère d'une séquence normale (la détection de mutations est très utile pour l'étude de troubles mentaux tels que l'autisme).

Dans le sens génétique, un marqueur représente n'importe quel caractère mesurable qui est relié à un locus de façon non ambiguë et qui permet de distinguer entre différentes variantes de ce locus. Le caractère essentiel d'un marqueur génétique est sa spécificité pour un locus chromosomique. Grâce aux marqueurs polymorphiques on peut distinguer quels allèles se trouvent chez un sujet et chez d'autres membres de la famille pour ainsi tracer la transmission d'un locus à travers les générations dans une famille.

Il y a plusieurs types de marqueurs génétiques dont les plus utilisés seront énoncés de façon abrégée:

2.3.1 Les RFLP, Polymorphismes de longueur des fragments de restriction

Les RFLP permettent de distinguer les molécules d'ADN des autres, en utilisant une technique de laboratoire qui coupe l'ADN selon la position du site de restriction (spécifique pour l'enzyme utilisé pour le détecter). L'ADN est coupé en quelques sites spécifiques comprenant, en général, un nombre pair de bases (4, 6 ou 8). Une enzyme ayant un site de reconnaissance à 6 bases coupe l'ADN toutes les 4 096 bases en moyenne. Le polymorphisme entre les individus se manifeste selon la distance entre deux sites de coupure de l'enzyme de restriction (site de restriction) ou bien pour la présence/absence du site de coupure, tel qu'illustré dans la Figure 2.3.

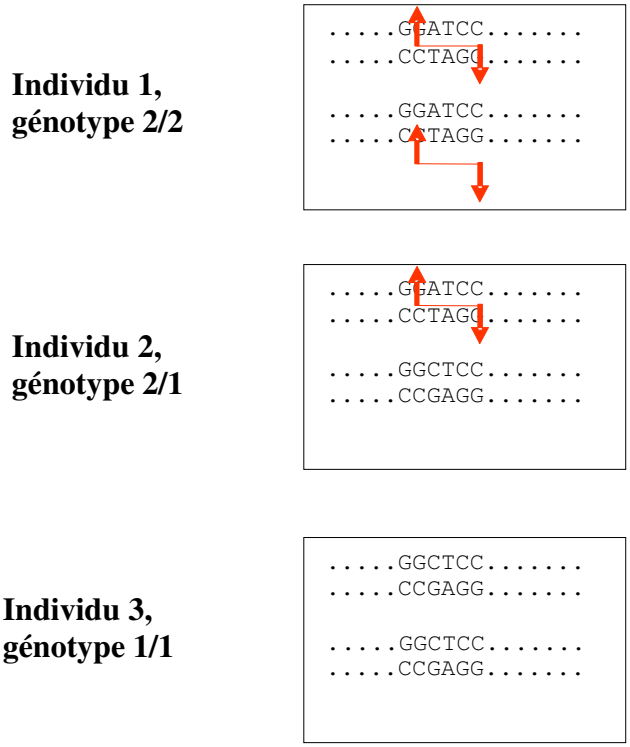


Figure 2.3 Un exemple de définition des RFLP, Polymorphismes de longueur des fragments de restriction

2.3.2 Les SNPs, polymorphismes à niveau d'un seul nucléotide

Les SNPs sont la source plus commune de variation entre les humaines et équivalent à des différences d'une seule base (A, C, G, T) entre les séquences génomiques. La Figure 2.4 montre des fragments de deux séquences, avec 8 SNPs. Il y a approximativement 3 millions de nucléotides variables dans le génome humain haploïde, soit une par chaque mille bases. Dans la Figure 4.1 on illustre l'utilisation d'un SNP pour vérifier la coségrégation d'une maladie et un gène spécifique.

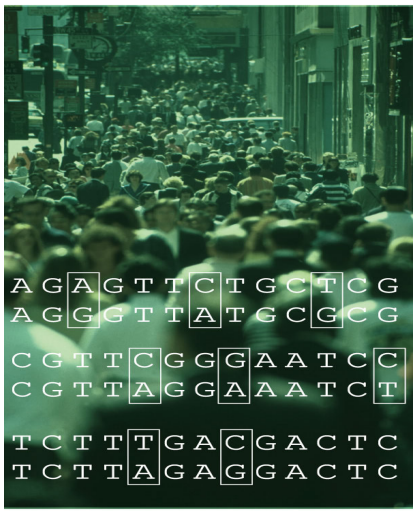


Figure 2.4 Les SNPs, polymorphismes à niveau d'un seul nucléotide tel qu'illustré dans NATURE (Chakravarti, 2001)

2.3.3 Les microsatellites

Les microsatellites sont aussi appelés *Short Tandem Repeats* (STR) et *Simple Sequence Repeats* (SSR). Les microsatellites sont des séries répétées d'un à quatre nucléotides répandus dans la séquence d'ADN. Comme le nombre de répétitions varie d'un individu à l'autre, les microsatellites sont des marqueurs très utilisés pour la cartographie d'immenses étendues du génome humain. La Figure 2.5 montre l'utilisation des microsatellites comme des marqueurs moléculaires: le modèle d'hybridation pour une famille de 6 enfants avec l'utilisation de quatre "allèles" microsatellites de tailles différentes (M' jusqu'à M''''). Le microsatellite (M'') est probablement lié à l'allèle P qui prédispose à la maladie et sa présence dans les individus atteints (le père et les enfants 1, 2, 5 et 6) permet de mettre en évidence la liaison entre l'allèle P et la maladie.

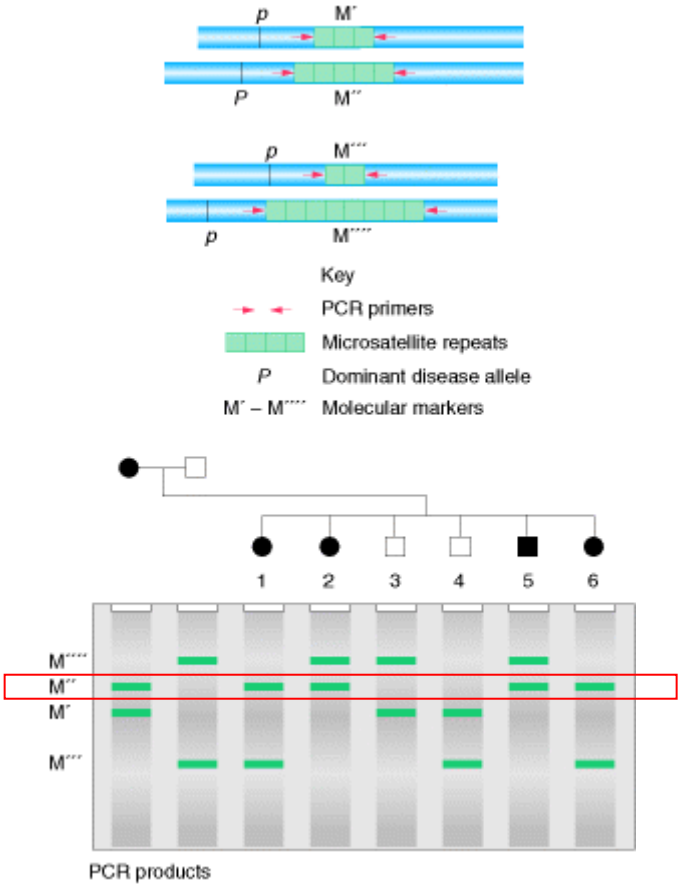


Figure 2.5 L'utilisation des microsatellites pour la cartographie d'un gène (Griffiths & Suzuki, 2002)

2.4 Mesure de la diversité génétique

Il est possible de décrire la composition génétique d'une population à partir des fréquences des différents allèles qui y sont présents. Une population est complètement décrite si l'on connaît la fréquence de chacune de ses catégories génétiques, la fréquence étant la proportion d'un trait sur l'ensemble de la population ciblée. La fréquence génotypique correspond à la proportion des individus porteurs d'un génotype et la fréquence allélique à la proportion des variantes correspondant à l'allèle en question dans la population. Par exemple, pour un caractère gouverné par deux allèles *A* et *a*, la structure d'une population de taille N est complètement connue si l'on connaît le nombre d'individus N_{AA}, N_{Aa} et N_{aa}, avec $N= N_{AA} + N_{Aa} + N_{aa}$, à partir desquels on calcule d'abord les fréquences relatives des trois génotypes, pour ainsi calculer les fréquences alléliques dans la population. Le nombre d'allèles A dans la population est $2*N_{AA}$ +

N_{Aa}. Les fréquences alléliques ***p*** et ***q*** des allèles A et a sont égales à : $p = (2*N_{AA} + N_{Aa})/2N$; $q = (2*N_{aa} + N_{Aa})/2N$; $p+q=1$.

Lorsque les croisements entre les individus d'une population sont aléatoires, c'est-à-dire, lorsque chaque descendant mâle a la même probabilité de se reproduire avec chacune des femelles, les fréquences génotypiques attendues dans la génération suivante peuvent être prédites à partir des fréquences alléliques des individus de la population parentale. On utilise alors le Tableau 2.1 pour déterminer les combinaisons possibles des génotypes des parents, en sachant que $p + q = 1$, ainsi:

$$(p+q)^2 = p^2 + 2pq + q^2 = 1 \qquad 2.1$$

Tableau 2.1 Tableau de Prunnet pour la détermination des fréquences alléliques des descendants de deux individus

Mère/Père	*A*	*a*
A	AA	Aa
	p^2	**pq**
a	Aa	aa
	pq	**q^2**

2.5 Principe d'équilibre d'Hardy-Weinberg

L'équation 2.1 exprimant la constitution génotypique de la descendance à partir de la fréquence allélique des parents est appelée la loi de Hardy-Weinberg. Son emploi implique que la diversité génétique d'une population se maintient et doit tendre vers un équilibre stable de la distribution génotypique. Grâce au modèle de Hardy-Weinberg on peut prédire la fréquence des génotypes à partir de la fréquence des allèles, la fréquence des allèles et des génotypes à partir de la fréquence des phénotypes et l'incidence des porteurs de maladies génétiques récessives rares. Si la population est dite en équilibre d'Hardy Weinberg alors la fréquence génotypique dans la population dépend seulement des fréquences des génotypes eux-mêmes. Si les fréquences génotypiques observées sur un marqueur s'éloignent des fréquences génotypiques attendues,

cela est expliqué soit par des erreurs de génotypage, soit parce que des forces extérieures s'appliquent ou encore parce que le mécanisme d'accouplement entre les individus n'est pas aléatoire.

2.6 Gènes et environnement : les maladies complexes

L'expression phénotypique d'un génotype est souvent due à l'interaction étroite entre des facteurs environnementaux (l'alimentation, le climat, les interactions avec d'autres espèces, les habitudes, etc.) et des facteurs génétiques transmissibles à la descendance, appelés interactions gène/environnement. Les maladies multifactorielles ou complexes sont toutes des pathologies dont l'origine réside à la fois dans les gènes et dans l'environnement. Le but de la recherche dans les maladies complexes est de mieux comprendre le processus de déclenchement des maladies, plutôt que la thérapie génique (consistant à remplacer un gène défaillant par un gène en bon état). La recherche génétique pour les maladies complexes porte donc sur des fluctuations et non sur des altérations comme pour les maladies monogéniques, dénommées aussi maladies mendéliennes (Gouyon, 2005). Une maladie monogénique est due à une altération dans la séquence d'un gène qui amène à la relation univoque : 1 mutation = 1 maladie.

Les études d'association entre les génotypes et les phénotypes se basent sur des méthodes statistiques pour trouver des déterminants génétiques dans les traits complexes comme l'asthme, l'hypertension ou la maladie d'Alzheimer, et sont basées sur deux types d'approche : le clonage positionnel ou les gènes candidats. Le clonage positionnel tente d'isoler des loci (ou gènes) de susceptibilité à la maladie à l'intérieur des régions chromosomiques identifiées par les analyses de liaison. L'approche par **gènes candidats** est basée sur l'étude des variations présentes à l'intérieur de gènes candidats. Ces derniers sont sélectionnés pour leur expression dans les tissus (suite à des études fonctionnelles) ou pour leur fonction connue et leur rôle biologique potentiellement important dans la physiologie de la maladie.

2.7 Le génome humain de référence au NCBI

Le Projet du Génome Humain a été un effort de 13 ans qui a livré en avril 2003 la première séquence de référence du génome humain. Pendant ce temps les centres de séquençage de partout dans le monde ont déposé des milliards de lettres de la séquence d'ADN des humains

dans GENBAN et ses bases de données associées, DDBJ et EMBL, et à partir de là les donnés ont été automatiquement mises à disposition des chercheurs. Aujourd'hui, la version définitive de la séquence du génome humain, les résultats des analyses issus du NCBI (Centre National pour l'Information biotechnologique) et leurs annotations sont disponibles pour consultation et téléchargement. Le génome humain de référence est composé de 24 chromosomes contentant 2.9 milliards de bases, couvrant approximativement 99% des positions dans le génome où se localisent les gènes. La séquence est exacte en moyenne au niveau d'une erreur à chaque 10000 bases. Des mises à jour sont introduites au fur et à mesure que des régions complexes sont affinées et que le petit nombre de lacunes entre les grandes étendues de séquences contiguës, ou "contigs", sont fermées.

L'outil *NCBI Map Viewer* permet de consulter sur une interface web et rechercher des séquences entières sur le génome de plusieurs organismes, produit des plans du génome en permettant d'obtenir de l'information à très haut niveau de détail sur n'importe quelle région d'intérêt. Elle centralise aussi des résultats des analyses de toutes sortes réalisées sur des marqueurs et répertoriées par plusieurs autres organismes. Nous pouvons y accéder à l'adresse: http://www.ncbi.nlm.nih.gov/mapview/. Nous utilisons cet outil pour valider la viabilité biologique des résultats obtenus dans les tests de liaison et d'association génétique, en raccordant les marqueurs qui sont significativement liés ou associés à la maladie aux gènes qui leur sont proches, et en conséquence à la fonction connue de ces gènes et leur rôle biologique potentiellement important dans la physiologie de la maladie.

CHAPITRE 3. LA STATISTIQUE EN BREF

Dans ce chapitre nous présentons un rappel des notions statistiques qui nous permettront de comprendre comment les énoncés sont modélisés et testés. Deux références principales sont cités: (Hines, et al., 2005) pour les notions basiques et (Sharma, 1996) pour l'analyse multidimensionnelle. Nous avons choisi de présenter séparément dans un chapitre postérieur le problème des tests multiples, qui découle naturellement de la statistique, mais qui pour nous représente le moteur principal des choix méthodologiques présentés dans ce mémoire.

L'étude décrite ici trouve ses outils et ses définitions dans une branche de la médecine connue comme épidémiologie génétique. L'épidémiologie génétique s'enrichit du rassemblement des méthodes génétiques et épidémiologiques indispensables pour élucider les causes et les facteurs d'une partie importante des maladies chroniques. Ainsi, elle se sert de la statistique principalement dans le but de décrire et d'analyser la variabilité. Les données statistiques dont on dispose proviennent en général d'un échantillon d'individus choisis au sein d'une population à laquelle on s'intéresse et proviennent de la mesure de plusieurs variables associées à ces individus. Un grand nombre de problèmes exigent, soit de faire une estimation, soit d'accepter ou rejeter un énoncé à propos d'un paramètre. L'énoncé est appelé une hypothèse et le processus de prise de décision est appelé un test d'hypothèse.

3.1 Les tests d'hypothèses

Une hypothèse est un énoncé sur la loi de probabilité à laquelle obéit une variable aléatoire. Le plus souvent, les hypothèses concernent un paramètre de cette loi. Les tests d'hypothèse sont effectués à partir des mesures prises dans un échantillon aléatoire de la population en question. Si les mesures sont compatibles avec l'hypothèse, celle-ci est dite vraie. La démarche de construction d'un test consiste d'abord à définir une hypothèse principale ou hypothèse nulle (H_0) en fonction d'un paramètre à tester. Une hypothèse alternative est une négation donnée de l'hypothèse nulle et sert à supporter la décision de rejeter l'hypothèse nulle. Par la suite une statistique est choisie en utilisant un estimateur du paramètre en question. Par exemple, Si H_0 est vraie une observation t de l'estimateur T doit être proche du paramètre, selon un seuil

critique en dessous duquel l'observation confirmera l'hypothèse alternative H_1 et rejettera donc l'hypothèse nulle.

Le seuil critique est choisi à partir des erreurs auxquelles les tests sont assujettis. Si l'on rejette une hypothèse nulle lorsqu'elle est vraie, on commet une erreur de type I. Si l'on accepte une hypothèse nulle lorsqu'elle est fausse, on commet une erreur de type II. Les probabilités de commettre ces types d'erreurs sont nommées α et β, et α est connu comme niveau du test ou seuil de signification (Hines, et al., 2005).

Selon (Hines, et al., 2005), la méthode la plus répandue dans les progiciels consiste à définir une probabilité critique (ou valeur p), probabilité à laquelle H_0 est conforme aux résultats obtenus sur l'échantillon. La valeur p est donc la probabilité que la statistique du test prenne une valeur au moins aussi grande que la valeur observée de cette statistique lorsque H_0 est vraie. Par exemple dans un test sur la moyenne d'une variable obéissant une loi gaussienne, la valeur p est relativement facile à calculer. Si Z_0 est la valeur calculée de la statistique, alors :

$$\text{valeur p} = \begin{cases} 2\left[1-\Phi(|Z_0|)\right] \textit{pour un test bilatéral}, \\ 1-\Phi(|Z_0|) \textit{ pour un test unilatéral à droite}, \\ \Phi(|Z_0|) \textit{ pour un test unilatéral à gauche.} \end{cases} \qquad 3.1$$

On utilise couramment la valeur p pour tester la validité de l'hypothèse nulle dans la plupart des problèmes en épidémiologie génétique. Par exemple pour décider si un marqueur est en liaison avec un phénotype donnée, un modèle de régression linéaire ou logistique est construit pour étudier la relation linéaire entre les variations phénotypiques et une certaine représentation de la transmission de l'information génétique des parents aux enfants. Ensuite l'hypothèse nulle que la pente de la régression est significativement différente de 0 est testée. Ne pas pouvoir rejeter l'hypothèse nulle équivaut a conclure qu'il n'y a pas de relation linéaire entre le trait et le marqueur. Le test est donc significatif si la valeur p du test est inférieur au seuil de signification $\alpha = 0.05$.

3.2 Les tests de permutation pour calculer la signification d'un test

Les tests de permutation sont utilisés pour déterminer si un effet observé tel que la différence de moyennes ou la corrélation entre deux variables est dû au hasard relié au choix de l'échantillon.

Sinon, on aura l'évidence que l'effet observé dans l'échantillon reflet un effet qui est présent dans la population. La façon de procéder consiste à:

1- Choisir une statistique quantifiant l'effet que l'on veut mesurer,
2- Construire la distribution nulle que cette statistique aurait eue si l'effet était présent dans la population,
3- Ranger la statistique observée dans la distribution nulle. Une valeur observée qui tombe dans le centre de la distribution peut être due au hasard. Une valeur à l'extrême de la distribution serait très rarement due au hasard, alors on a une évidence que quelque chose différent du hasard est en jeu.

Les méthodes de rééchantillonnage, dont les permutations font partie permettent d'utiliser les données observées exhaustivement dans le but de faire des inférences. Nous proposons plusieurs analyses qui utilisent les méthodes de rééchantillonnage et nous en parlerons d'avantage dans les chapitres suivants.

3.3 Méthodes de réduction de dimensionnalité pour des variables quantitatives

La réduction de dimensionnalité d'un groupe de variables quantitatives permette d'obtenir un plus petit ensemble de variables indépendantes qui encapsulent la variance totale des variables originales, en se basant principalement sur les mesures de covariance ou corrélation entre les variables originales. Autre les statistiques descriptives qui nous permettent d'obtenir une représentation généraliste des données, deux autres méthodes ont été choisies parce qu'ils tiennent compte de la corrélation entre les phénotypes et modélisent la variabilité: l'analyse de Composantes principales et l'analyse factorielle.

3.3.1 Analyse des Composantes Principales (ACP)

Le but de l'analyse en composantes principales est de réduire l'ensemble des variables à quelques composantes non corrélées entre elles qui expliquent la plus grande proportion de la variance possible entre les données. En assumant qu'existent p variables, on s'intéresse à fournir un ensemble de p combinaisons linéaires

$$\begin{aligned}
\xi_1 &= w_{11}x_1 + w_{12}x_2 + \ldots + w_{1p}x_p \\
\xi_2 &= w_{21}x_1 + w_{22}x_2 + \ldots + w_{2p}x_p \\
&\vdots \\
\xi_p &= w_{p1}x_1 + w_{p2}x_2 + \ldots + w_{pp}x_p,
\end{aligned} \qquad 3.2$$

tel que ξ_1, ξ_2, …, ξ_p sont les p composantes principales et w_{ij} est le poids de la j^e variable sur la i^e composante principale. Les poids w_{ij} sont estimés de manière que:

- La première composante principale, ξ_1, compte pour la variance maximale des variables; La deuxième composante principale, ξ_2, compte pour la variabilité maximale des variables qui n'a pas été expliqué par la première composante, et ainsi de suite.
- $w_{i1}^{\ 2} + w_{i2}^{\ 2} + \ldots + w_{ip}^{\ 2} = 1 \quad i = 1,\ldots, p$
- $w_{i1}w_{j1} + w_{i2}w_{j2}^{\ 2} + \ldots + w_{ip}w_{jp}^{\ 2} = 0 \quad \forall i \neq j$.

La solution du problème peut être obtenue par des différentes techniques, la recherche des valeurs et vecteurs propres de la matrice de covariance est la plus utilisée. Toutefois, l'analyse en composantes principales peut être accomplie en recherchant la décomposition de valeurs régulières de la matrice de données ou par la décomposition spectrale de la matrice de covariance (Sharma, 1996).

3.3.2 Analyse factorielle

Le but de l'analyse de factorielle est d'identifier les facteurs sous-jacents qui expliquent la corrélation entre les variables $x_1, x_2, \ldots, x_p$. En assumant on a p variables, on s'intéresse à fournir un ensemble de p combinaisons linéaires dans un modèle de m facteurs :

$$\begin{aligned}
x_1 &= \lambda_{11}\xi_1 + \lambda_{12}\xi_2 + \ldots + \lambda_{1m}\xi_m + \varepsilon_1 \\
x_2 &= \lambda_{21}\xi_1 + \lambda_{22}\xi_2 + \ldots + \lambda_{2m}\xi_m + \varepsilon_2 \\
&\vdots \\
x_p &= \lambda_{p1}\xi_1 + \lambda_{p2}\xi_2 + \ldots + \lambda_{pm}\xi_m + \varepsilon_p \quad .
\end{aligned} \qquad 3.3$$

λ_{pm} est le facteur de pondération (*loading*) de la p^e variable dans le m^e facteur et ε_p est le facteur unique de la p^e variable. La corrélation entre les variables est expliquée par les *m* facteurs communs. Normalement on assume que le nombre de facteurs communs est plus petit que le nombre d'indicateurs *p*. Dans d'autres mots, la corrélation entre les p indicateurs est due aux *m* (*m*<*p*) facteurs. Si les facteurs ne sont pas corrélés entre eux, le modèle est dénommé orthogonal, au cas contraire le modèle est nommé oblique.

Les équations précédentes peuvent être représentées sous forme matricielle pour produire l'équation principale de l'analyse factorielle :

$$x = \Lambda\xi + \varepsilon, \qquad 3.4$$

où $\boldsymbol{x}$ est un vecteur de taille p x 1 contenant les variables, $\boldsymbol{\Lambda}$ est une matrice *p x m* du schéma des facteurs de pondération et $\boldsymbol{\varepsilon}$ est une vecteur de taille p x 1 contenant les facteurs uniques. On assume que les facteurs ne sont pas corrélés avec les composantes $\boldsymbol{\varepsilon}$ et les variables ont été d'abord standardisées (elles ont toutes une moyenne égale à zéro et sa variance est égale à 1). La matrice de corrélation **R** des indicateurs est donnée par :

$$\begin{aligned} \mathrm{E}(xx') &= \mathrm{E}[(\Lambda\xi + \varepsilon)(\Lambda\xi + \varepsilon)'] \\ &= \mathrm{E}[(\Lambda\xi + \varepsilon)(\xi'\Lambda' + \varepsilon')] \\ &= \mathrm{E}[(\Lambda\xi\xi'\Lambda)] + \mathrm{E}[(\varepsilon\varepsilon')], \\ R &= \Lambda\Phi\Lambda' + \Psi \end{aligned} \qquad 3.5$$

Φ étant la matrice de corrélation des facteurs et **Ψ** une matrice diagonale contenant les variances uniques. Les communalités (la corrélation expliquée par les facteurs) sont dans la diagonale de la matrice **R - Ψ** . Les éléments hors de la diagonale sont les corrélations entre les indicateurs. Les matrices **Φ**, **Ψ** et **Λ** sont les matrices à estimer étant donné la matrice de corrélation **R**.

Pour un modèle orthogonal, l'équation antérieure peut être réécrite :

$$R = \Lambda\Lambda' + \Psi, \text{ ou bien,}$$

$$\Lambda\Lambda = R - \Psi \qquad 3.6$$

Le côté droit de l'équation équivaut à la matrice de corrélation avec les communalités dans la diagonale. Les estimations des facteurs de pondération sont obtenues en calculant la structure des valeurs et vecteurs propres de la matrice **R - Ψ**. Néanmoins, l'estimation de **Ψ** est faite en résolvant l'équation **:**

$$\Psi = R - \Lambda\Lambda'. \qquad 3.7$$

Ainsi, la solution de l'équation 3.6 requiert la solution de l'équation 3.7 et vice-versa. Cette circularité amène au problème d'estimation des communalités (Sharma, 1996).

Plusieurs méthodes sont utilisées pour réaliser l'analyse factorielle. La fonction *factanal* du logiciel R (R-Development-Core-Team, 2007) utilise la méthode de vraisemblance maximale pour calculer la matrice des facteurs de pondération et la matrice de variance des facteurs uniques. Ce modèle présume que les unicités sont distribuées selon une distribution normale multidimensionnelle.

Les nouvelles variables, appelées scores sont aussi estimées, soit par régression multiple, soit par la méthode des carrés minimums de Bartlett. L'objectif est d'estimer les scores non observés f, tels que :

$$\hat{f} = X'\Phi. \qquad 3.8$$

La méthode de Thomson fait une régression dans la population de la fonction *f* (*x*) inconnue, tel que

$$\hat{f} = \Lambda'\Psi^{-1}X, \quad 3.9$$

et substitue par après les estimations des quantités à droite de l'équation ci-dessus. La méthode de Bartlett minimise la somme des carrés des erreurs standard sur le choix de f, étant donné la valeur estimée de Λ.

3.3.3 Représentation graphique des composantes principales et des facteurs

Les valeurs propres représentent l'importance des nouveaux caractères et servent à la fois à définir la proportion de la variance totale des variables originales exprimées par les nouvelles composantes. Le graphique des valeurs propres en fonction du nombre de composantes principales est utilisé dans le but de décider combien de composantes expliquent suffisamment bien (généralement dans le sens de la variance cumulée) les variables d'entrée. Une règle de

sélection consiste à choisir les composantes dont la valeur propre est plus grande qu'un. Cela s'explique du fait que toutes les variables d'entrée ont été standardisées et que, en conséquence, la variance de chacune de ces variables est égale à 1. On pourrait donc dire que chaque composante choisie doit expliquer plus que ce qu'une variable d'entrée est capable d'expliquer elle-même. La Figure 1.6 montre les valeurs propres des composantes obtenues à partir de nos données phénotypiques. 6 phénotypes intermédiaires mesurés 5 fois selon un protocole clinique visant à montrer les variations des phénotypes reliés à l'hypertension dans diverses positions orthostatiques. Selon la règle de "valeur propre plus grande que 1", on choisirait dans l'exemple de la Figure 3.1 les quatre premières composantes, expliquant ensemble plus du 80% de la variance totale.

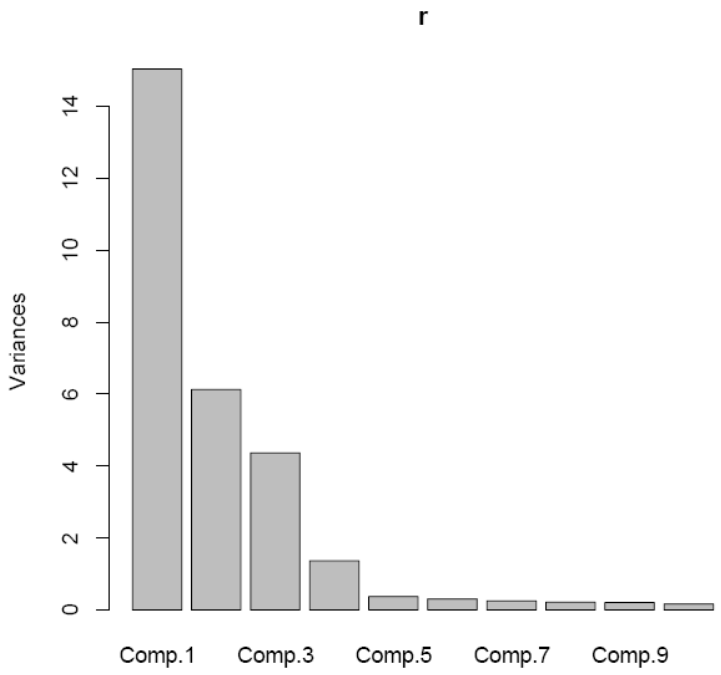


Figure 3.1 Représentation des valeurs propres dans un exemple d'application de l'analyse de composantes principales

Un autre type de figure représentative de l'ACP et l'analyse factorielle consiste à projeter les variables d'entrée dans le plan des composantes principales. Cela équivaut à dessiner les facteurs de pondération (corrélation entre les variables originales et les nouvelles composantes) projetés dans un cercle de corrélation de rayon 1 centré à zéro. La Figure 3.2 montre le cercle des corrélations pour le même exemple de la Figure 3.1. Les variables sont projetées sur le plan des composantes 1 et 2. Clairement un voit un effet de regroupement des variables

correspondant à des mesures répétées. La magnitude des facteurs de pondération pour les phénotypes TPR et SV est plus grande sur la composante 1. Celle des facteurs de pondération pour la pression systolique (SYS) sur la composante 2, le pouls (HR) sur la composante 3 et la pression diastolique (DIA) sur la composante 4. Sur tous les plans les variables HR et SV sont orthogonales par rapport à SYS et DIA, ainsi que TPR par rapport à SYS et DIA. Ce fait nous permet de dire que dans le contexte de l'analyse les variables HR, SV et TPR sont indépendants des pressions artérielles SYS et DIA.

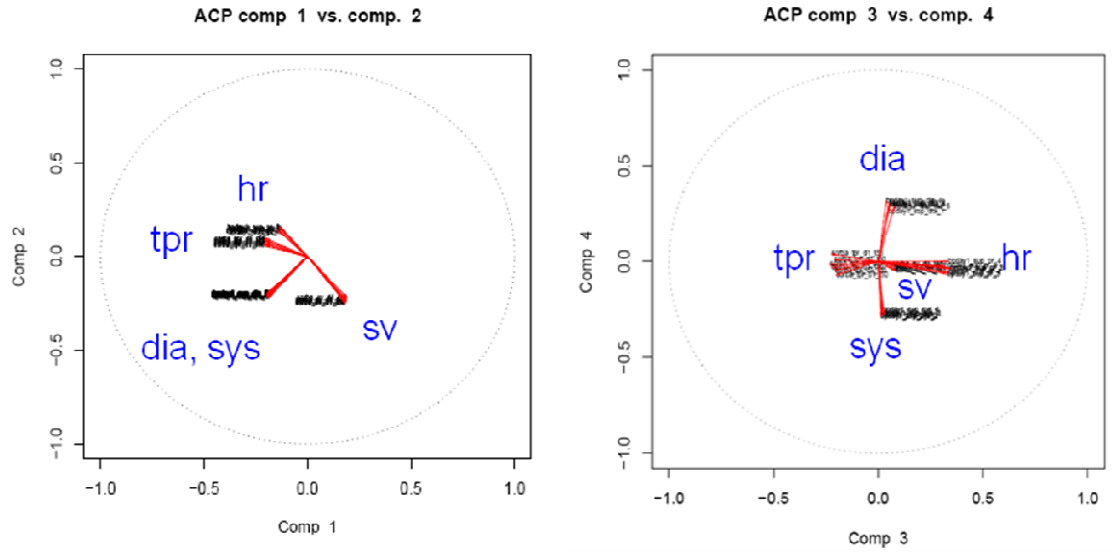


Figure 3.2 Représentation des facteurs de pondération (*loadings)* dans un exemple d'application de l'analyse de composantes principales.

CHAPITRE 4. TESTS DE LIAISON ET D'ASSOCIATION GÉNÉTIQUE

Une stratégie commune pour trouver des déterminants génétiques dans les maladies complexes consiste à utiliser l'analyse de liaison sur un grand ensemble de marqueurs génétiques (cet ensemble pouvant s'étendre jusqu'à l'analyse de tout le génome), suivi d'une analyse d'association pour raffiner la cartographie de la variation génétique dans les régions qui sont liées. Dans ce chapitre nous présentons les deux tests utilisés pour les analyses auxquelles nous sommes engagés: le test de liaison basé sur les pairs de germains et le test de déséquilibre de transmission pour tester l'association génétique familiale.

Le principe dans la plupart des approches de cartographie génétique est que l'allèle qui prédispose à une maladie se transmettra de génération en génération accompagné d'autres variantes qui sont très fortement liées à une position donnée dans le génome. La recombinaison génétique est un principe important affectant la liaison: plus un gène et un marqueur sont proches l'un de l'autre, moins il y a de chance qu'il y ait de la recombinaison pouvant les séparer. La recombinaison est utilisée afin de déterminer la distance génétique entre le gène et le marqueur et se mesure en centimorgans, étant 1 centimorgan (cM) équivalant à 1% de recombinaison. Les études de liaison examinent directement la transmission intergénérationnelle des deux phénomènes : le phénotype, résultant dans une maladie, et les allèles d'une famille, en cherchant des corrélations qui suggèrent que le marqueur génétique est lié à un locus causal. Dans l'analyse de liaison paramétrique, la transmission de la maladie et du marqueur est évaluée selon le modèle de transmission en utilisant des techniques de vraisemblance maximale sur des modèles statistiques des familles étendues. Dans l'analyse non paramétrique, la mesure d'excès de partage d'allèles est prise juste sur les individus atteints, ce qui évite le besoin de l'établissement d'un modèle de maladie (Balding, 2006). Cette méthode, connue sous le nom d'analyse de sib-pair (pairs de germains), teste l'hypothèse de liaison en estimant la proportion d'allèles partagés identiques par descendance entre frères et sœurs. Sous l'hypothèse nulle de non-liaison, la transmission des allèles d'un marqueur génétique donné, des parents à leurs enfants, se fait au hasard et la moyenne des proportions d'allèles partagés par toutes les paires possibles de germains affectés dans une étude est égale à 0.5. Un excès de partage d'allèles parmi les paires d'affectées (c'est-à-dire une proportion moyenne > 0,5) indique une distorsion de la distribution aléatoire due à une liaison génétique entre le marqueur et la maladie (Ailhaud & Institut national de la santé et de la recherche médicale, 2000).

Dans la Figure 4.1 l'étude consiste à vérifier la co-ségrégation d'un phénotype humain (une maladie par exemple) et un marqueur spécifique, un SNP dans l'exemple. Si les individus qui ont hérité la maladie héritent aussi le SNP (marqueur génétique), alors le gène qui cause la maladie et le SNP ont tendance à être proches dans le chromosome, tel qu'il est montré dans la figure. Pour vérifier que la liaison est significative statistiquement, plusieurs individus doivent être examinés. Occasionnellement le SNP sera séparé du gène causant la maladie par des événements de recombinaison pendant la formation de l'ovule ou du spermatozoïde, ce qui s'est passé dans la paire de chromosomes à droite. Le gène qui cause la maladie est hérité avec le SNP à partir de la mère atteinte dans 75% de sa descendance. Si la même corrélation est observée dans d'autres familles examinées, le gène qui cause la maladie est cartographié dans le chromosome proche du SNP. Un SNP plus éloigné sur le chromosome ou dans un autre chromosome loin du gène ciblé sera co-ségrégué 50% des fois, ce qui équivaut dans les tests de liaison à l'hypothèse nulle (Alberts, 2002).

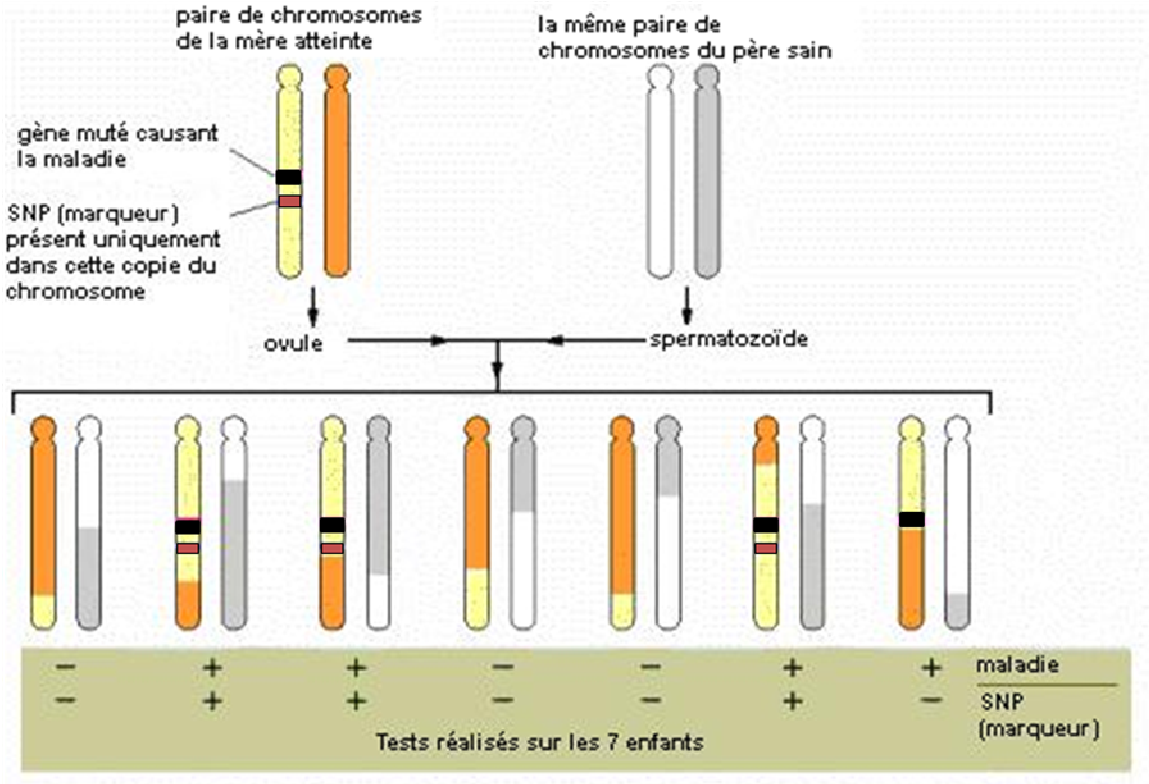


Figure 4.1 Un exemple d'utilisation de l'analyse de liaison génétique en utilisant des marqueurs physiques sur l'ADN (SNPs) pour la cartographie d'un gène (Alberts, 2002).

Un grand avantage des tests de liaison est que l'information est combinée à travers les familles de façon à ce que l'évidence d'un rôle causal d'un locus puisse s'additionner même si d'autres variantes se séparent de ce locus dans d'autres familles différentes. L'analyse de liaison est appropriée quand plusieurs variantes rares d'un locus contribuent au risque d'une maladie. Les analyses de liaison requièrent des grandes familles ou un grand ensemble de familles pour acquérir un niveau de signification et de résolution importante (Balding, 2006).

Les analyses d'association génétique répondent à une question qui semble être simple à poser, soit: quels allèles sont surreprésentés dans des individus malades par rapport aux individus sains? Le but des analyses d'association est d'identifier les patterns des polymorphismes qui varient systématiquement entre les individus en consonance avec des différents niveaux de la maladie et qui peuvent désormais représenter l'effet d'un allèle de risque ou d'un allèle protecteur. Un problème présent dans les analyses d'association est celui de la taille du génome. Il est tellement long que les patterns suggérant un polymorphisme causal ont pu apparaître plutôt par hasard. Pour distinguer entre les associations réelles et celles aberrantes, on peut soit utiliser des restrictions données par les méthodes statistiques, soit ne choisir qu'un ensemble de polymorphismes pouvant être générés par des variantes génétiques causales, étant donné notre connaissance sur l'histoire génétique des humains et les événements évolutionnaires tels que les mutations et les recombinaisons (Balding, 2006).

4.1 Logiciels pour les analyses de liaisons et associations familiales

Le choix de logiciels pour les analyses de liaison et d'association familiale est fait en général sur la base des expériences des chercheurs et sur les limitations des logiciels. Au laboratoire du Dr. Pavel Hamet au Centre de recherche du CHUM on donne la priorité aux logiciels supportant des familles étendues et des grandes familles (puisque la majorité d'études sont faits sur plusieurs familles d'une même population), en plus du support offert par les constructeurs et de la correspondance du modèle par rapport à l'information existante. Le module SIBPAL (*sibling pair analysis*) du logiciel SAGE (S.A.G.E., 2002) est utilisé pour l'analyse de liaison. Le module SIBPAL calcule des tests des moyennes, des tests des proportions et résout un modèle basé sur des régressions linéaires combinant:

- La différence des moyennes carrées pondérées des traits entre les pairs de germains (deux enfants des mêmes parents) et
- la somme carrée corrigée par la moyenne des traits,

étant la variable dépendante la mesure de la proportion d'allèles identiques par descendance (*identical by descent*, IBD), c'est-à-dire, provenant d'un même allèle parental, tel que montré dans la Figure 4.2. Le modèle est connu comme le modèle de Haseman-Elston (HE) (Haseman & Elston, 1972). Pour comprendre le modèle, faisons une révision de la notation basique fournie par le logiciel S.A.G.E. (S.A.G.E., 2002). Une famille nucléaire est un ensemble de deux individus qui se sont reproduits et leurs enfants; ces enfants forment des couples de germains complets. Les cousins et cousines forment avec ces enfants des couples de mi-germains.

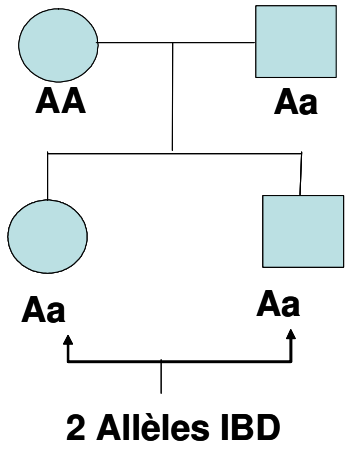


Figure 4.2 Allèles identiques par descendance. Deux allèles sont identiques par descendance s'ils sont copies du même allèle des parents.

Soit le nombre de familles dans l'analyse *P*;
soit $n_p = 1,2\ldots \Sigma_{pnp} = n$ le nombre de couples de germains (complètes et mi-germains) dans la p^e famille, où n est le nombre total de couples de germains.

Conditionnellement à l'information des marqueurs génétiques disponible dans une position particulière dans le génome; soit $\hat{f}_{1j}$ la probabilité de partager 1 allèle identique par

descendance (IBD) pour le j[e] pair et $\hat{f}_{2j}$ la probabilité de partager 2 allèles IBD pour le j[e] couple. À noter que $\hat{f}_{2j} = 0$ pour les mi-germains.

Soit

$$\pi = (1 + 2w_1)/4 \quad et \quad \hat{\pi}_j = \hat{f}_{2j} + w_1 \hat{f}_{1j} \quad où \quad 0 \le w_1 \le 0.5\,, \qquad 4.1$$

pour le j[e] pair de germains. La valeur par défaut de w_1 est 0.5.

Le modèle pour l'analyse de liaison est donné par une régression de la forme :

$$\mathrm{y} = \beta_0 + \sum_m a_m \hat{\pi}_m + \sum_m d_m \hat{f}_{2m} + \sum_m c_k f(z_k) + \varepsilon \quad , \quad 4.2$$

où

y est la variable dépendante qui est une fonction de la variable phénotypique,

β_0 est l'ordonnée à l'origine.

π_m est donné par l'équation **,** 4.1,

c_k est un facteur de nuisance équivalant à l'effet des covariables sur le modèle

ε est l'erreur résiduelle.

Dans un échantillon aléatoire lorsque la valeur de w_1 est 0.5, a_m est la variance génétique additive et d_m est la variance génétique dominante due au marqueur. Étant **β** le vecteur des paramètres pour une telle modèle linéaire. L'estimateur de **β** est calculé par la méthode des moindres carrés comme suit :

$$b = (A^{'} W^{-1} A)^{-1} A^{'} W^{-1} y\,; \qquad 4.3$$

et celle de la variance résiduelle est:

$$s^2 = \frac{y^{'} W^{-1} (y - Ab)}{n - m}\,, \qquad 4.4$$

étant m le nombre de paramètres estimés, n le nombre de couples de germains et *y* un vecteur de taille nx1 des variables dépendantes (les variables phénotypiques). *W* est une matrice de poids de taille n x n pour *y* et A est une matrice de taille n x m qui contient dans chaque colonne

un paramètre du modèle. La matrice de poids *W* peut être soit une matrice de corrélation, soit la matrice de variance résiduelle Σ de y, son choix dépendant de la méthode utilisée pour générer la variable dépendante *y*.

L'hypothèse nulle de non-liaison entre le marqueur et le caractère est vérifiée selon un test t de la statistique b_i/v_i^2, où b_i est l'i^e élément des paramètres à estimer selon l'équation 4.3 et v_i^2 est l'estimé de la variance, qui est le produit de la i^e diagonale de $(A' W^{-1} A)^{-1}$ et s^2, calculé selon l'équation 4.4. Étant donné que l'indépendance des couples ne peut pas être assurée, le module SIBPAL utilise la méthode itérative de l'équation estimante généralisée (GEE de l'anglais *generalized estimating equations*). D'abord, toutes les corrélations sont initialisées à zéro pour obtenir les résidus; par après, les corrélations des résidus de l'itération antérieure sont utilisées pour mettre à jour la matrice des poids W et pour l'estimation des paramètres b_i et s_i^2. Les itérations s'arrêtent si la variation de la corrélation résiduelle d'une itération à l'autre est plus petite qu'un seuil prédéfini.

La combinaison des méthodes de liaison et d'association est mise à disposition grâce aux analyses d'association familiale. L'association entre les SNPs et les phénotypes est testé au moyen du test TDT (*transmission disequilibrium test*) disponible dans le logiciel FBAT (de l'anglais *Family Based Association Tests*) (Laird, Horvath, & Xu, 2000) reposent sur la statistique générale U, basée elle-même dans la combinaison linéaire des génotypes et des caractères. En considérant un nombre *n* de familles nucléaires (composées chacun des parents biologiques et des enfants) indépendantes, l'hypothèse nulle devient donc:

H_{01} : il n'y a ni liaison ni association entre les marqueurs et la maladie.

La statistique du test pour la famille i est donnée par l'équation:

$$U = S - E[S], \qquad S = \sum_{ij} T_{ij} X_{ij}, \quad 4.5$$

où X_{ij} définit une fonction du génotype de la j^e génération de la famille *i* dans le locus qui est testé. La statistique dépend entre autres du modèle génétique en question. T_{ij} est une fonction du caractère et peut être affecté par d'autres paramètres inconnus. Dans le cas particulier des traits quantitatifs, la codification de T_{ij} devient $T_{ij} = Y_{ij} - \mu_{ij}$. Y_{ij} est la valeur observée du caractère et μ_{ij} une valeur de compensation (*offset* en anglais), qui est utilisée pour donner une

connotation particulière à l'analyse (soit pour la définition de covariables, pour l'analyse des contrastes de la transmission ou d'autres qui seront précisés ci-dessous).

Alternativement, les poids des contributions des individus atteints ou non atteints peuvent être altérés par l'utilisation de l'*offset.* Par exemple, par l'utilisation de la valeur μ_{ij} = 0.5 dans le cas d'un caractère dichotomique (de valeur 0 où 1), T_{ij} devient -0.5 pour le sujet atteint et 0.5 pour le sujet sain. Les individus dans ce cas reçoivent le même poids mais avec des différents signes, et la statistique expose un contraste des transmissions des individus atteints versus les non atteints. Dans le cas des caractères quantitatifs, la valeur de μ_{ij} recommandé est la moyenne de l'échantillon. La valeur T_{ij} peut être aussi ajustée pour une covariable z_{ij} en convertissant l'équation $T_{ij} = Y_{ij} - \mu_{ij}$ en $T_{ij} = Y_{ij} - \beta_0 - \beta_1 z_{ij}$.

La statistique de l'équation 4.5 permet aussi de définir les modes de transmission génétique entre dominante, récessive et additive. Le
Tableau 4.1 montre la codification de X_{ij}, correspondant au génotype de l'i^{e} individu appartenant à la j^{e} famille selon le mode de transmission choisi (additif, récessif, dominant) et selon le génotype de l'individu. Le modèle additif exprime bien le comportement de la transmission génétique, même si le véritable modèle génétique n'est pas additif.

Tableau 4.1 Codification du génotype dans les analyses d'association familiale avec FBAT pour un modèle biallélique.

Mode de transmission	***Génotype***		
	AA	Aa	aa
Dominante pour A	1	1	0
Récessive pour A	1	0	0
Additive pour A	2	1	0

La statistique de FBAT se base sur la distribution des génotypes de la descendance conditionnés sur les phénotypes et sur les génotypes des parents (géniteurs). Si les génotypes des parents sont inconnus, la statistique est conditionnée sur la distribution de la descendance. Étant donné que la distribution des génotypes de la descendance peut être calculée en utilisant

les lois de ségrégation de Mendel, l'approche de FBAT est considérée comme étant robuste envers le manque de spécification du modèle.

En utilisant la distribution des génotypes de la descendance, c'est-à-dire en fixant T_{ij} et en considérant X_{ij} comme aléatoire, *V = Var (U) = Var (S)* peut être calculé sous l'hypothèse nulle et utilisée pour standardiser U. Si X_{ij} est une valeur scalaire du génotype de l'individu, la statistique de l'échantillon sera:

$$Z = \frac{U}{\sqrt{V}}, \qquad 4.6$$

qui suit approximativement une loi gaussienne de paramètres N(0,1).

Si X_{ij} est un vecteur alors

$$\chi^2 = U^{'} V^{-1} U, \qquad 4.7$$

suit approximativement une loi χ^2 avec le nombre de degrés de liberté égal au rang de V.

Étant donné que nous savons préalablement que le marqueur est lié au phénotype, l'hypothèse nulle qui est plus appropriée est connue comme H_{02} : pas d'association en présence de liaison. Le test de cette hypothèse tient compte de la liaison entre les génotypes des descendants d'une même famille, et du fait que lorsque la liaison existe, les transmissions génétiques d'un parent à ses enfants ne sont pas indépendantes. Selon Laird (Laird, et al., 2000), FBAT permet d'utiliser le même test sous l'hypothèse de "non-association et non-liaison" et d'ajuster la variance par un estimateur de la corrélation entre les génotypes et par la présence de plusieurs familles nucléaires au sein d'une famille étendue. Dans l'étude ici présenté le mode de transmission additif est privilégié parce qu'il exprime bien le comportement de la transmission génétique, même si le véritable modèle génétique n'est pas additif. Nous l'avons choisi par rapport à l'alternative de tester tous les modèles et accepter l'ajustement par test multiples qui en découle.

CHAPITRE 5. MÉTHODES DE CORRECTION ACTUELLES

Dans ce chapitre nous abordons le problème des tests multiples, en présentant les méthodes de correction qui on été développés pour contrôler le taux de fausses découvertes lorsque plusieurs tests sont testés. Nous présentons quelques méthodes de correction pour tests multiples pour ensuite indiquer pourquoi ces méthodes ne s'appliquent pas aux problèmes étudies.

5.1.1 Le problème des tests multiples

Le problème des tests multiples se réfère à toute situation dans laquelle une suite de tests statistiques sont évalués (Westfall & Young, 1993) . La raison qui justifie la correction par tests multiples est qu'en appliquant plusieurs tests d'hypothèse à la fois on augmente la probabilité de déclarer des résultats faussement significatifs, c'est-à-dire, associations statistiquement significatives qui ne le sont pas.

Par exemple, dans une étude particulière les chercheurs testent quelques hypothèses et trouvent un résultat significatif pour l'une d'elles avec une valeur p = 0.005. Donc, les chercheurs déclarent que l'effet est statistiquement significatif. Cette valeur p est interprétée comme suit: lorsque la relation de causalité n'existe pas pour l'effet testé (l'hypothèse nulle est vraie), il y a une probabilité de 0.05% d'observer un résultat aussi extrême que le résultat observé. Les chercheurs exécutent un procédé qui calcule une valeur p ajustée *p-adj*= 0.15, qui n'est pas statistiquement significative. Cette nouvelle valeur p tient compte des multiples tests et peut être interprétée comme suit: lorsque la relation de causalité n'existe pas pour *n'importe lequel* des effets testés, il y a une probabilité de 15% que *quelque part dans l'expérience* on observe un résultat aussi extrême que le résultat observé de 0.005 (Westfall & Young, 1993).

En utilisant un langage formel dans un problème de tests multiples (PTM) un ensemble de *m* hypothèses, disons H_{01}, H_{02}, ..., H_{0m}, sont testées simultanément, avec la condition que la probabilité d'une ou de plusieurs erreurs de type I est inférieur ou égal à une constante α. Pour chaque hypothèse H_i une statistique du test est disponible avec sa valeur p correspondante P_i. Soient les hypothèses $\{H_{01}, H_{02}, ..., H_{0m_0}\}$ celles pour lesquelles l'hypothèse nulle est vraie et $\{H_{11}, H_{12}, ..., H_{1m_0}\}$ celles fausses. Si la statistique du test est continue, P_{0i} est distribué selon

une distribution uniforme dans l'intervalle [0,1]. La distribution marginale de chaque P_{1i} est inconnue, mais si les tests ne sont pas biaisés, elle est stochastiquement plus petite que la distribution des P_{0i}. Soient *R* le nombre d'hypothèses rejetées par un problème de tests multiples, *V* le nombre d'hypothèses nulles rejetées erronément et *S* le nombre de faux positifs rejetés. D'entre ces trois variables aléatoires, seule R peut être observée. L'ensemble de valeurs p, $\mathbf{p}=(\mathbf{p_0}, \mathbf{p_1})$, et $r=v+s$ obtenues dans un problème de tests multiples sont les événements des variables aléatoires définies. (Yekutieli & Benjamini, 1999)

En suivant les termes définis ci-dessus, le nombre attendu de rejets erronés, ainsi appelé le taux d'erreur par famille (PFE de l'anglais *per family error-rate*) est $E_P V(p)$. Le risque global d'erreur, appelé *"Family Wise Error Rate"* (FWER), est la probabilité de rejeter au moins une hypothèse qui est vraie, FWE = $Pr_{\mathbf{p}}$ (Yekutieli & Benjamini, 1999). Pour un test d'hypothèse simple, le seuil de signification α peut être défini comme "rejeter lorsque la valeur $p <= \alpha$". Dans un PTM on définit la signification comme: Pour chaque $p \in [0,1]$, rejeter H_0 si $p_i <= p$. Le PFE d'une telle procédure est $E_P V(p)= m_0.p$. Rejeter toutes les hypothèses nulles équivaut à rejeter l'hypothèse correspondante à la valeur p minimale observée de toutes les hypothèses nulles vraies, le FWE de cette procédure est Pr_{P0} *s.*

5.2 Contrôler le taux d'erreur au sein d'une famille de tests

Le taux d'erreurs que nous voulons contrôler est le FWER. Cela consiste à s'assurer que la probabilité de rejeter au moins une hypothèse nulle vraie est au plus α. Il existe deux types de FWER: le taux d'erreur lorsqu'un sous ensemble des hypothèses nulles est vrai (FWEP, sous l'hypothèse nulle partielle) et le taux d'erreur lorsque toutes les hypothèses nulles sont vraies (FWEC, sous l'hypothèse nulle complète). Parce que l'on doit contrôler que n'importe quel sous-ensemble d'hypothèses nulles est vrai, le contrôle de FWEP est plus rigoureux que le contrôle de FWEC et on l'appelle contrôle forte du FWER. En contrepartie, contrôler le FWEC devient contrôler faiblement le FWER (Westfall & Young, 1993).

5.3 La correction de Bonferroni

L'ajustement de Bonferroni est une procédure générique des PTM dont le seuil de signification devient $p = \alpha/m$, ce qui équivaut à ajuster chaque valeur p en le multipliant par m. Donc, une hypothèse nulle sera rejetée par la méthode de Bonferroni si sa valeur p ajustée *p-adj* est

inférieur à α. La procédure contrôle fortement la FWER. C'est démontré que lorsque les tests sont indépendants et qu'ils sont suffisamment puissants, aucune correction qui contrôle le FWER ne peu produire de meilleures valeurs p corrigés. Par contre, lorsque les statistiques des tests sont corrélées la correction de Bonferroni peut être conservatrice. Et dans les tests d'association plusieurs sources de corrélation existent: les marqueurs proches sont corrélés à cause du déséquilibre de liaison, les phénotypes intermédiaires de la pression artérielle sont corrélés, les mesures répétées des phénotypes sont aussi corrélés. (Brunelle, 2008) a démontré que sur le total de SNPs autosomiques génotypés dans la population étudiée (54524), à un seuil de corrélation (la mesure la plus utilisée du déséquilibre de liaison) de 0.8, plus d'un tiers de SNPs sont corrélés. L'analyse factorielle sur l'ensemble de phénotypes testés nous a permis d'obtenir 6 variables à partir de 55, en conservant plus de 80% de la variabilité des donnés. Les phénotypes sont donc corrélés. Nous pouvons prévoir que les statistiques de tests seront corrélées et que la correction de Bonferroni est conservatrice pour notre ensemble.

5.4 Le taux de fausses découvertes (FDR)

Les procédures basées sur l'espérance du taux de faux positifs (FDR de l'anglais *False Discovery Rate*) contrôlent le taux d'erreur de type I en fonction de la proportion V / (V + S) de faux positifs parmi les hypothèses rejetées. Pour la méthode d'ajustement basé sur le taux de faux positifs (FDR), l'ajustement de la valeur p consiste à comparer chaque valeur p ordonnée $p_{(i)}$ avec $\alpha.i\ /m$. Soit k= max s, si un tel k existe, rejeter $H_{0(1)}$, ... , $H_{0(k)}$. (Yekutieli & Benjamini, 1999). Ces procédures sont particulièrement attrayants pour les problèmes à grande échelle, comparativement aux taux d'erreur traditionnels basés sur le nombre de faux positifs V (par exemple, la FWER = P_r (V> 0)), car ils ne croissent de manière exponentielle avec le nombre M d'hypothèses testées. Et malgré cela elles restent conservatrices parce qu'elles tiennent compte de la corrélation entre les tests que pour quelques structures de corrélation restreintes.

Selon (Brunelle, 2008), les applications pour lesquelles on applique des tests qui contrôlent le taux de faux découvertes pour un ensemble de tests d'association ou de liaison génétique corrélés sont incertaines pour le moment. Le taux de fausses découvertes dépend aussi de la puissance du test, qui dépende à sa fois de la taille de l'échantillon, de la fréquence de l'allèle que l'on teste, de la magnitude de l'effet que l'on veut mesurer et, naturellement, du type de test que l'on applique. Or, la proportion attendue de fausses découvertes varie d'une expérience à l'autre.

5.5 Les méthodes bayésiennes

Alternativement à l'ajustement par tests multiples, on pourrait ajouter une connaissance à priori pour ajuster les niveaux de signification en utilisant des méthodes statistiques bayésiennes. Ainsi, si l'on sait à priori qu'un effet particulier est réel, la méthode bayésienne permettra que l'effet soit déclaré réel même si l'évidence statistique est faible. Il s'agit donc de combiner les valeurs p observés, la puissance du test et la probabilité à priori que l'association soit vraie. Les études de puissance tiennent compte de la taille de l'échantillon, de la fréquence de l'allèle que l'on teste, de la magnitude de l'effet que l'on veut mesurer et, naturellement, du type de test que l'on applique. (Westfall & Young, 1993) disent qu'on devrait s'abstenir d'ajuster par tests multiples si une connaissance à priori est apportée. Sous l'approche bayésienne, on n'est pas pénalisé pour analyser les donnés exhaustivement parce que la probabilité à priori d'une association ou d'une liaison ne devrait pas être affectée par les tests que le chercheur réalise (Balding, 2006). La difficulté principal est justement l'établissement des probabilités à priori. Un marqueur pour lequel plusieurs liaisons ou associations ont été déjà publiés aurait une probabilité à priori plus haute qu'un autre marqueur quelconque. Et par la suite d'autres suppositions peuvent être considérées. Il en résulte que dans différents études différentes questions sont posés et que les calculs de puissance sont réalisés sur des suppositions différentes, ce qui produira des résultats très différents et dans quelques cas, fort opposés.

5.6 Les méthodes de rééchantillonnage

On peut disposer d'autres méthodes que l'ajustement postérieur de la valeur *p* pour contrer le problème des tests multiples. Étant donné que les limitations computationnelles deviennent de plus en plus surmontables, on pourrait agir tel qu'un bon chercheur le ferait dans la pratique: il répéterait l'expérience jusqu'à ce qu'il ait épuisé toutes les combinaisons possibles des paramètres de son essai. Les méthodes de rééchantillonnage permettent d'utiliser les données observées exhaustivement dans le but de faire des inférences. Les variables observées sont donc réassignées aléatoirement aux individus en étude et les statistiques sur ces nouveaux arrangements sont recalculées des milliers de fois. La valeur de la statistique observée est donc considérée inusuelle si elle est inusuelle par rapport à la distribution calculée par rééchantillonnage. Lorsqu'une seule hypothèse est testée par rééchantillonnage, la valeur p est calculée sans faire de suppositions sur la distribution des données, ce qui donne des tests plus

puissants que ceux basés sur des distributions paramétriques lorsque leurs suppositions ne sont pas respectées. Lorsque plusieurs hypothèses sont testées, la valeur p calculée tient compte de tous les tests qui ont été effectuées en préservant la corrélation entre les tests, ce qui leur rendre plus puissant que les méthodes d'ajustement qui n'évaluent que le nombre de tests effectuées (parce que en général ils considèrent que les tests sont indépendantes) (Westfall & Young, 1993).

Une des erreurs communes associées aux méthodes de rééchantillonnage est l'erreur de simulation (Brunelle, 2008). L'erreur de simulation représente le fait qu'un nombre fini d'échantillons ne représente pas la population réel. Ainsi, on doit contrôler que le nombre de permutations effectuées soit suffisant. Pour ce faire, l'erreur standard de la valeur p ajustée doit être calculé par rapport au nombre de permutations réalisées. Si l'intervalle de confiance a 95% après N permutations comprend le seuil α=0.05, les simulations devraient se poursuivre. L'erreur standard doit être calculée ainsi:

$$es=\sqrt{\frac{p_{adj}(1-p_{adj})}{N}},$$

Etant p_{adj} la valeur p empirique calculé è partir de la méthode de rééchantillonnage, et N le nombre de permutations effectués. Les générateurs de nombres pseudo-aléatoires introduisent aussi des erreurs dans les ajustements des valeurs p par rééchantillonnage. Dans la section de validations nous décrivons les validations qui ont été faites par rapport aux tels générateurs.

En résumé, les méthodes d'ajustement qui n'évaluent que le nombre de tests effectués à la manière de la correction de Bonferroni s'avèrent conservateurs pour notre étude, car ils ne tiennent pas compte de la corrélation entre les unités de test. La méthode que contrôle le taux de fausses découvertes fournit en générale des ajustements moins conservateurs que les méthodes FWER, mais pour l'instant nous ne sommes pas en condition d'établir un modèle de la structure de corrélation des tests ni de la puissance des tests de liaison. Pour l'établissement des probabilités à priori dans les tests de liaison, nous préférons donner la même chance à tous les SNPs étant donné que tous sont des bons candidats pour être liés aux phénotypes testés. Nous avons choisi d'agir en mode exploratoire. Les méthodes de rééchantillonnage sont, en effet, un choix exceptionnelle, et nous sommes prêts à payer le prix pour l'utilisation de ces méthodes: le temps d'exécution et la quantité de test effectués. Nous mettons l'accent sur les

optimisations algorithmiques et les structures de données nécessaires pour accélérer les tests, ce qui rend possible d'avoir des valeurs p ajustées dans un délai raisonnable.

Nous proposons plusieurs analyses qui utilisent les méthodes de rééchantillonnage. D'abord nous utilisons un test de permutation pour confirmer que les statistiques du test de liaison varient selon les stimuli environnementaux externes. Ensuite nous utilisons une méthode de permutation pour quantifier la signification de l'effet des tests physiologiques par rapport aux marqueurs disponibles, la généalogie et les phénotypes mesurés. Finalement nous proposons plusieurs méthodes de réduction de dimensionnalité des phénotypes pour tester l'association entre les phénotypes intermédiaires de l'hypertension et l'ensemble des marqueurs significativement liés détectés dans la premier partie de l'étude. Des simulations par rééchantillonnage des phénotypes sont effectués pour vérifier le control de l'erreur de type I dans les tests d'association et pour ajuster seuil de signification par rapport aux données observées. La méthodologie proposée sera exposée dans le chapitre suivant.

CHAPITRE 6. RAPPEL DU PROBLÈME ET MÉTHODES PROPOSÉES

Dans ce chapitre nous faisons un rappel du problème en présentant, à la manière des textes scientifiques en génétique, la caractérisation de l'hypertension, suivie de la présentation de la population étudiée. Nous montrons aussi comment la qualité des génotypes a été vérifié et nous introduisons le concept de liaison dynamique. Par la suite, nous décrivons la méthodologie proposée en détaille.

6.1 L'hypertension: une maladie complexe

La pression artérielle, un caractère complexe définissant l'hypertension (pression artérielle haute), est un caractère quantitatif qui varie continuellement. Les causes de l'augmentation de la pression artérielle sont inconnues dans 95% des cas. Sur le 5% restant, quelques processus connus contribuant à l'hypertension ont pu être identifiés et les sujets atteints ont été catégorisés comme atteints de "hypertension secondaire". Les individus pour lesquels on ne peut pas identifier le mécanisme sont décrits comme atteints d'hypertension essentielle, décrite pour Schröder (Cowley, 2006) comme "une maladie simple dont la cause est inconnue ou un ensemble de maladies séparées ayant en commun une origine et plusieurs effets secondaires, dont l'ensemble de maladies sont étiologiquement, pathologiquement et génétiquement semblables à plusieurs égards ". L'hypertension essentielle est la maladie cardiovasculaire plus commune avec une prévalence[3] de presque 26 % à travers du monde. Elle est un facteur de risque pour les accidents vasculaires cérébraux, les maladies du cœur et pour l'insuffisance rénale terminale (Cowley, 2006). L'hypertension est reconnue comme une maladie polygénique, dont quelques variantes monogéniques constituent seulement le 1% des cas (Hamet, et al., 2005).

Les traits complexes sont très souvent influencés par des facteurs qui ne sont pas d'intérêt direct. La pression artérielle peut être déterminée par d'autres phénotypes intermédiaires qui sont reliés à la fonction vasculaire, cardiaque et neuroendocrine. La Figure 6.1 montre que la pression artérielle est définie par le débit cardiaque (*cardiac output*) et la résistance vasculaire totale (*Total Peripheral Resistance*). Le débit cardiaque est défini par un mélange complexe de

[3] La prévalence est la proportion d'individus malades présents à un moment donné dans une population par rapport à la population en entier.

relations entre le volume extracellulaire des fluides, le volume du sang et la résistance des artères et des veines envers le flux sanguin autour du système circulatoire. Le rein contrôle autant le volume de fluide extracellulaire que la pression artérielle. Les facteurs neuroendocrines ont aussi une influence importante autant sur la fonction des reins que sur la fonction vasculaire d'autant que quelques mécanismes de rétroaction qui fonctionnent comme des espèces de senseurs de la pression artérielle produisant des signaux qui se retransmettent et servent ainsi à contrôler la fonction neuroendocrine, vasculaire et des reins (Cowley, 2006).

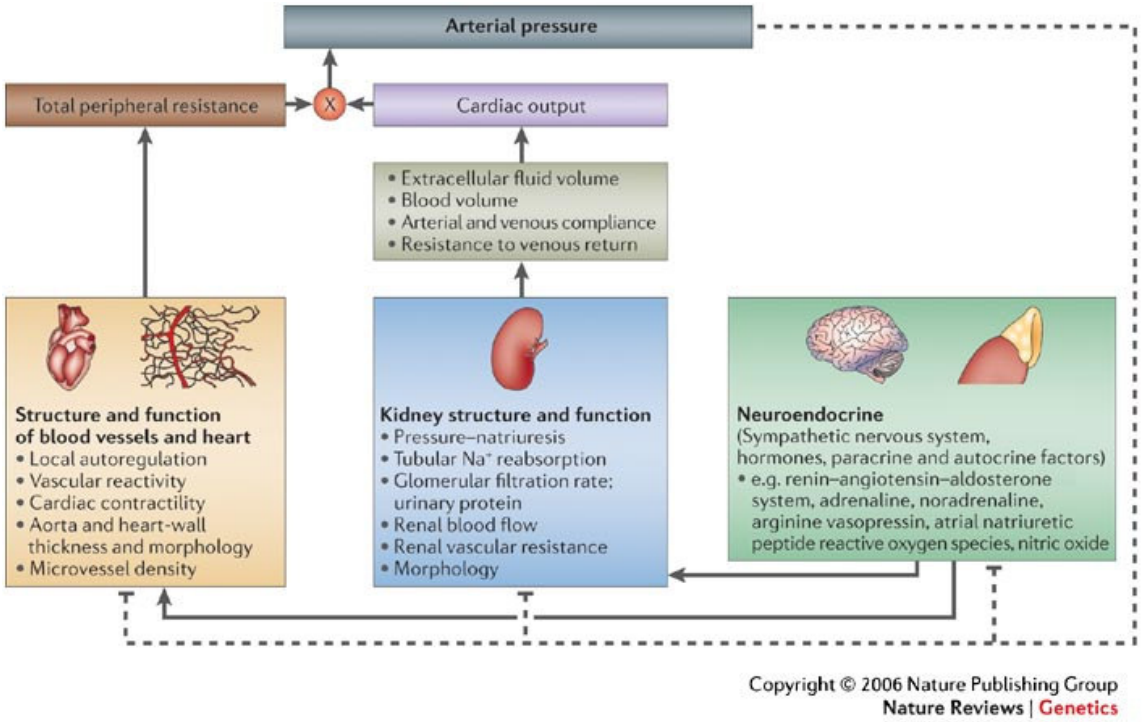


Figure 6.1 Traits intermédiaires définissant la pression artérielle selon (Cowley, 2006)

D'autres facteurs tels que le sexe, l'âge, le poids et les facteurs environnementaux tels que la consommation d'alcool ou de tabac peuvent affecter l'état d'une maladie et théoriquement devraient être tous mesurés et connus si l'on veut entreprendre l'analyse du trait complexe. Généralement on traite ces dernières en ajustant les valeurs des phénotypes selon une analyse de régression, à savoir, en les incluant comme covariables. Ces ajustements sont sensibles à l'erreur et au manque de spécifications et les valeurs p des tests d'association ou liaison dépendent des variables d'ajustement choisies. En conséquence, l'interprétation des tests devient difficile et la reproductibilité des études est mise en question, spécialement si les

nouvelles études n'ont pas pu mesurer et ajuster le modèle selon les mêmes variables utilisées dans les études publiées.

Les avances sur la définition de la base génétique de la susceptibilité à l'hypertension sont retardées parce que la pression artérielle est un caractère complexe et polygénique influencé par des diverses variantes, des interactions gène-gène et par l'environnement. Dans quelques cas, il peut y avoir des multiples variantes, chacune avec un petit effet; dans d'autres cas, juste quelques variantes pourraient être responsables. La Figure 6.2 montre les locus (emplacements sur les chromosomes) liés à l'hypertension dans l'ensemble du génome humain, tel que rapporté dans la littérature. Les barres entourant les chromosomes représentent la localisation des locus significatifs. Les barres plus foncés à l'intérieur des chromosomes montrent des emplacements sur les chromosomes qui n'ont jamais été rapportés (Cowley, 2006).

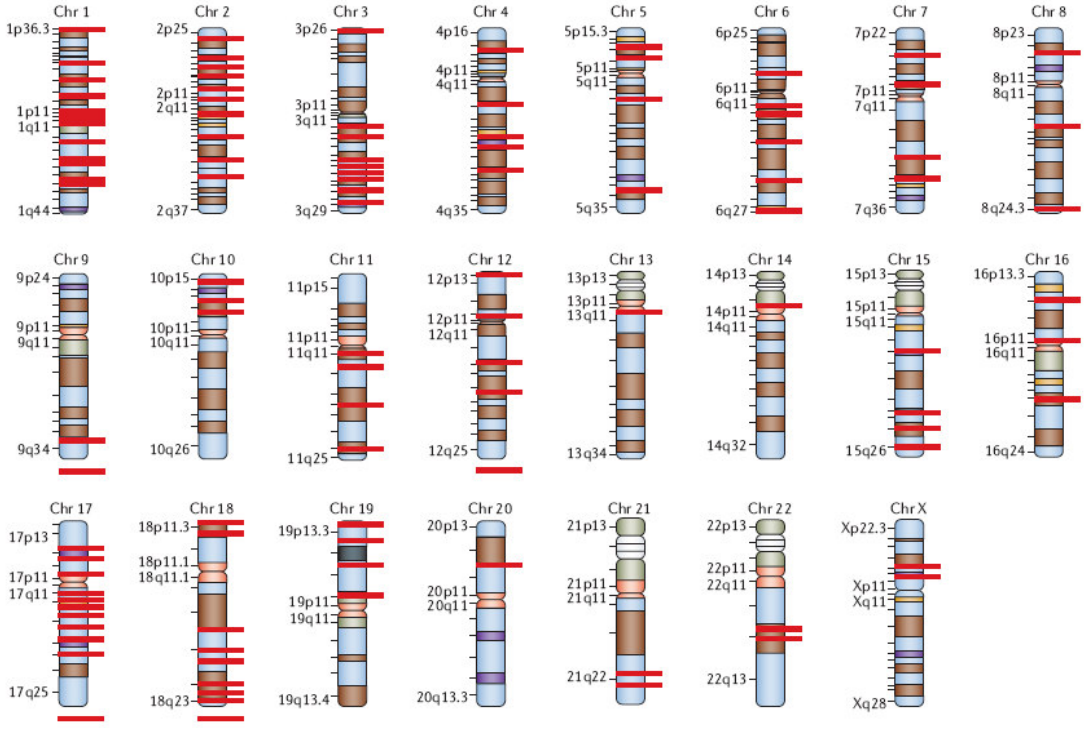


Figure 6.2 Locus quantitatifs de l'hypertension dans le génome humain (Cowley, 2006)

Dans la recherche systématique sur l'ADN des facteurs de risque des maladies complexes, on examine à la fois plusieurs points sur le génome pour savoir si un facteur de risque se trouve à proximité de l'un d'eux. Dans la mesure où le signal inspecté est très faible, on s'expose à deux risques : soit d'interpréter comme positif un signal qui n'est que le fruit du hasard, soit de ne rien

détecter alors qu'il y avait quelque chose. C'est ce qu'on définit comme des faux positifs ou des faux négatifs. L'incidence de faux positifs est proportionnelle au nombre d'essais et au seuil de signification des tests. Or, l'application de méthodes unidimensionnelles (un phénotype, un marqueur) aux traits complexes manque de puissance. Les petits effets marginaux et les effets d'interaction entre multiples facteurs ne sont pas détectés par le test unidimensionnel. En conséquence, l'étude des maladies multifactorielles impose la mise à point de procédures qui évitent les tests multiples (Gouyon, 2005). Nous ajoutons au problème de la multiplicité due à la présence d'un nombre important de marqueurs génétiques le fait d'utiliser des mesures répétées de plusieurs phénotypes. Ce fait pourrait augmenter la force des signaux de liaison d'autant qu'augmenter le nombre de tests effectués, et en conséquence, réduire la puissance du test en affectant le nombre de faux positifs.

La réduction de dimensionnalité des phénotypes est proposée dans le but d'obtenir une meilleure représentation des phénotypes par rapport aux critères suivants:

- les différences entre les phénotypes qui sont dues aux tests physiologiques,
- la variabilité due aux mesures répétées, représentant la dynamique des phénotypes,
- la variabilité due à la corrélation entre les phénotypes.

Cette réduction est basée sur l'application de trois méthodes : l'analyse en composantes principales, l'analyse factorielle et les statistiques descriptives (la moyenne). Les mesures répétées des multiples phénotypes sont soumises à des méthodes qui tiennent compte parfois de la corrélation entre les variables, parfois de la variabilité totale ou bien par des mesures qui met en évidence les différences mesurables par rapport aux tests physiologiques.

6.2 Population étudiée, phénotypage et génotypage

La population cible dans cette étude est celle du Saguenay-Lac-St-Jean (SLSJ), reconnue comme une population à effet fondateur. Cette population offre plusieurs avantages par rapport aux populations générales, entre autres parce que les individus dans les populations à effet fondateur partagent un environnement homogène (ce qui limite l'influence de l'environnement sur les paramètres héréditaires dans l'expression clinique des maladies); les possibilités de cartographie sont plus grandes parce que les intervalles en déséquilibre de liaison sont plus longs et parce que l'on connaît amplement sa généalogie (Hamet, et al., 2005). Les registres

généalogiques de la population cible ont été compilés dès 1608 et ont été informatisés dans le registre de population BALSAC (Bouchard, Roy, Casgrain, & Hubert, 1989). Par rapport à d'autres populations, où les mélanges ethniques ont contribué à hétérogénéiser le bagage génétique, les individus dans des populations dites à effet fondateur conservent dans leur génome des blocs ancestraux de taille considérable qui sont communs à tous les individus, contribuant à la conservation d'une certaine homogénéité.

Des individus appartenant à une cohorte de familles du Saguenay-Lac-Saint-Jean ont été soumis à certaines manœuvres orthostatiques et quelques traits reliés à la pression artérielle ont été mesurés en utilisant l'impédance cardiaque, tel que montré dans laFigure 6.3. Le phénotypage a été réalisé par du personnel entraîné suivant des procédés standardisés. Le jour 0 de l'expérience consistait à la prise d'échantillons sanguins pour l'extraction de L' ADN et pour la mesure d'autres paramètres physiologiques. Le jour 1, un cathéter intraveineux à été inséré dans les patients pour la prise de tests de lipides, glucose et concentration d'insuline ainsi qu'un test de posture représenté dans la Figure 6.3. Pendant 60 minutes les individus restaient couchés et ensuite des périodes de mesure de 10 minutes en position debout et 20 minutes en position assise se sont suivies. Des variables cardiaques telles que le pouls (HR), la résistance vasculaire totale (TPR) et la pression sanguine, ont été mesurés par pléthysmographie d'impédance.

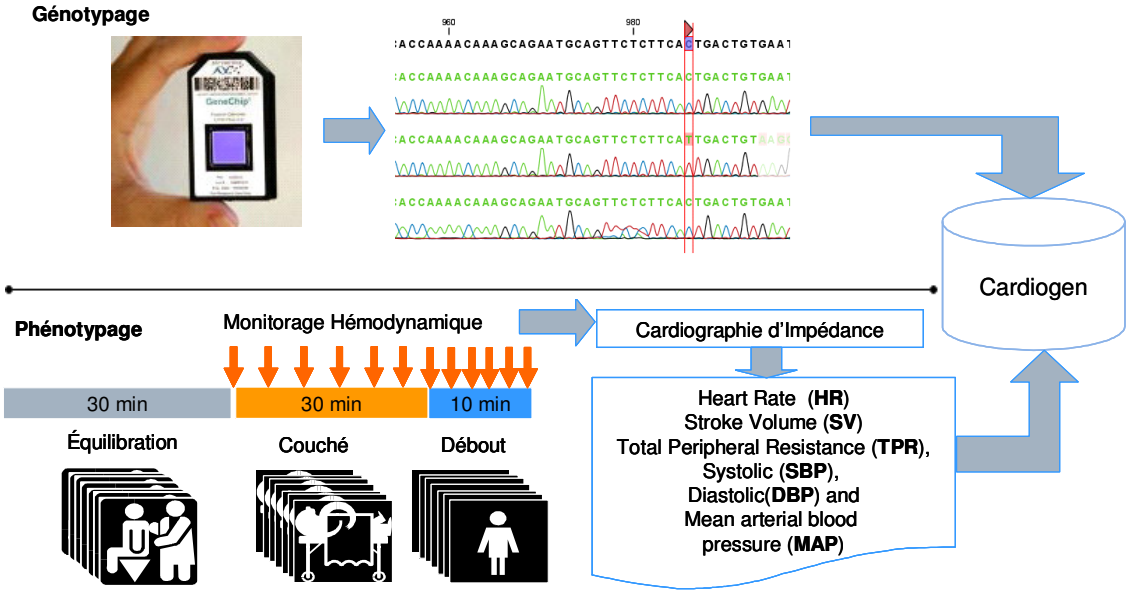


Figure 6.3 Représentation des étapes de génotypage et phénotypage dans les études de liaison dynamique.

L'information génétique et phénotypique a été par la suite organisée dans la base de données CARDIOGEN selon un modèle de base de données relationnelle. Pour l'extraction de l'information de la base de données, les chercheurs se servent d'une application cliente du même nom qui permet d'extraire des fichiers formatés selon les spécifications des divers logiciels utilisés pour l'analyse génétique.

6.3 Sélection des données et vérification de la qualité des génotypes

La base de donnés ciblée dans la présente étude est celle du système CARDIOGEN implanté au Centre de Recherche du CHUM à l'Hôtel Dieu. Les familles ont été recrutées à partir d'une population dans le Saguenay–Lac-St-Jean, une région de la province du Québec au Canada. Les familles ont été choisies selon la présence d'au moins une paire de germains atteintes d'hypertension ou dyslipidémie. Les familles ont entre 2 et 11 enfants et proviennent de 76 familles. La condition pour l'inclusion des paires de germains était l'existence d'hypertension essentielle (pression sanguine systolique (SBP) > 140 mm Hg et/ou pression sanguine diastolique (DBP) > 90 dans deux périodes de temps différents ou la consommation de médication anti hypertensive), dyslipidémie (cholestérol plasmatique ≥ 5.2 mmol/l et/ou cholestérol HDL ≥ 0.9 mmol/l ou la consommation de médicaments pour la diminution du taux de lipide), indice masse corporelle (DBP) < 35 kg/m^2, âge entre 18 et 55 ans, et origine catholique canadien français (Hamet, et al., 2005). Les participants ont été génotypés en utilisant les pouces Affymetrix Xba 50k. Ce pouce contient environ 58000 SNPs repartis sur les 23 chromosomes. La totalité des sujets pour lesquels un phénotypage complet a été effectué ont été génotypés.

Dans l'approche par **gènes candidats** les variations présentes à l'intérieur de gènes candidats sont vérifiées selon que les polymorphismes soient à l'intérieur ou en proximité des gènes identifiés comme associés à un trait par des études antérieures. Des gènes impliqués dans la fonction vasculaire, la fonction autonomique, la fonction cardiaque, le pouls, la fonction des reins, la résistance à l'insuline et le syndrome métabolique seraient de bons candidats pour l'hypertension. Une fois que le gène est sélectionné, le ou les marqueurs disponibles (les polymorphismes d'un seul nucléotide ou SNPs), qui serviront de marqueur génétique, sont choisis. Dans la présente étude, ces SNPs ont été choisis partir de la base de données du Centre National d'Information sur les Biotechnologies (NCBI), en utilisant les mots clés "vascular

function" (fonction vasculaire), "autonomic function" (fonction autonome), "cardiac function" (fonction cardiaque), "heart rate" (pouls), "kidney function" (fonction rénal) "hypertension" (hypertension), "insulin resistance" (résistance à l'insuline) et metabolic syndrome" (syndrome métabolique), ainsi que dans la base de données OMIM en utilisant le mont clé hypertension ou hypotension. La liste a été aussi additionnée des gènes candidats publiés par publiées par Dobson, Jr. et al, Halushka et al à (Halushka, et al., 1999), Okuda et al, (Okuda, et al., 2002) et Iwai et al (Iwai, et al., 2004). La plupart des marqueurs choisis sont positionnés sur des gènes. Les SNPs avoisinantes ont été aussi filtrées selon qu'ils étaient en déséquilibre de liaison avec ceux localisés à l'intérieur des gènes candidats ou dans la base de données du projet HapMap pour les Caucasiques (populations CEU, originaires de l'Europe du Nord et de l'Ouest). Ensuite les SNPs ont été sélectionnés si leurs fréquences alléliques étaient en équilibre de Hardy-Weinberg ($p>0.01$), et si la fréquence de l'allèle mineur était plus grande ou égale à 5%. Les tests des fréquences alléliques et de déséquilibre de liaison ont été calculés en utilisant un test exact décrit en Wigginton et al (Wigginton, Cutler, & Abecasis, 2005).

6.4 Liaison dynamique

L'analyse de liaison dynamique définit l'analyse de liaison exécutée sur des phénotypes dont la valeur varie à travers le temps sous l'influence de traitements physiologiques. Les tests de liaison issus de l'approche de Haseman et Elston fournis par SIBPAL (S.A.G.E., 2002) sont appliqués à des ensembles d'individus soumis à des conditions de stress physiologique et produisent des courtes séries de temps de valeurs t. Un test rapide de permutation est utilisé pour tester l'hypothèse que les changements orthostatiques ont un effet significatif sur le degré de liaison génétique et pour trouver les SNPs qui causent cet effet. La correction de la valeur p de signification de la liaison sur chaque période est aussi calculée par le rééchantillonnage des génotypes et est rapportée dans la section des simulations. Sur cette dernière approche, nous avons confirmé les résultats du test de simulation rapide en utilisant une approche qui tient compte du fait que l'on teste la liaison en présence de plusieurs marqueurs génétiques et des mesures répétées des phénotypes.

Le problème de liaison dynamique ne s'est jamais posé comme tel dans la littérature avant nous. Un problème qui lui ressemble a été proposé dans l'atelier de génétique humaine GAW13 en 2003 dans le cadre d'une étude d'association sur les phénotypes longitudinaux, c'est-à-dire, ceux dont la valeur est mesurée pendant des intervalles de la vie des individus. L'atelier avait

comme base des données de l'étude Framingham Heart, une initiative très reconnue dans le domaine cardiovasculaire crée dans le but d'élucider les risques reliés aux maladies cardiovasculaires. Des solutions bases sur la réduction unidimensionnelle ont été proposées, et les créateurs de FBAT ont répondu en produisant un test (FBAT-PC) qui applique la réduction de dimensionnalité par une combinaison de l'analyse de composantes principales et l'héritabilité des phénotypes (Lange, et al., 2004). Malheureusement, la structure de nos familles est tellement complexe et dans quelques cas les familles sont si grandes que le programme n'est pas capable de réaliser les tests. Nous avons donc réalisé nous-mêmes les réductions et analysé les effets des réductions par simulations.

En exploitant la disponibilité d'une source importante d'information des phénotypes dynamiques liés à la variation de la pression artérielle dans une population génétiquement homogène, l'objectif du présent travail est d'identifier et quantifier l'effet des tests physiologiques sur la liaison génétique. Cette quantification devrait apporter à l'équipe de génomique prédictive du Centre de Recherche du CHUM des pistes sur les processus complexes qui déclenchent l'hypertension.

6.4.1 Tests de permutation pour la liaison dynamique

Dans une expérience typique d'analyse de liaison génétique des milliers de SNPs et phénotypes sont inclus simultanément. Lorsque les phénotypes proviennent d'individus soumis à différentes conditions physiologiques, il s'avère nécessaire de tester les différences significatives dans le degré de liaison qui sont dues à ces conditions. L'approche naturelle consiste à traiter chaque couple phénotype - SNP comme un essai indépendant et appliquer un test individuellement. Dans le cas particulier de la quantification de l'effet du test physiologique, on applique automatiquement un test de corrélation entre le degré de liaison et une représentation de la condition expérimentale. L'incidence de faux positifs (SNPs qui sont étiquetés comme liés dynamiquement qui ne le sont pas) est proportionnelle au nombre d'essais et au seuil de signification des valeurs p.

Les corrections par tests multiples ajustent les valeurs p issues de multiples tests statistiques pour corriger selon l'occurrence de faux positives. La pratique habituelle dans les problèmes de tests multiples est d'utiliser la correction de Bonferroni par nombre de comparaisons, en assumant l'indépendance des essais (un essai étant défini par la série de temps des valeurs t pour un SNP et un phénotype donné). Cette supposition n'est pas observée, parce qu'une

certaine corrélation de la liaison est attendue, soit parce que nous traitons des SNPs reliés à gènes candidats qui pourraient être liés ensemble, soit parce que les phénotypes sont corrélés, donc les statistiques du test de liaison sont corrélées.

Les tests de permutation font partie des méthodes de rééchantillonnage. Ils permettent de quantifier l'incertitude d'une hypothèse en calculant des statistiques et des tests de signification. Ils requièrent moins de suppositions que les tests traditionnels et donnent généralement des réponses plus précises. Les tests de signification sont utilisés pour établir si un effet observé (dans notre cas la corrélation entre deux séries de temps) est raisonnablement dû au hasard introduit par la sélection de l'échantillon. Sinon, ils donnent l'évidence que l'effet observé dans l'échantillon reflète un effet qui est présent dans la population (Moore, McCabe, Duckworth, & Sclove, 2003).

Le principe de localisation de SNPs liés dynamiquement calcule une statistique (par exemple la corrélation) entre les valeurs t issues des tests de liaison génétique et une courbe en escalier représentant la condition expérimentale (en assignant un nombre entier à chaque position). En supposant que l'hypothèse nulle est vraie, c'est-à-dire, que la condition expérimentale a un effet nul sur les statistiques de liaison, n'importe quel réarrangement des observations produira une valeur de la statistique mesuré très petite. Le test de permutation utilise le rééchantillonnage (réarrangement des valeurs t) pour construire une estimation empirique que la statistique aurait si l'effet était présent dans la population. Dans chaque itération la séquence de l'observation est réordonnée et la corrélation entre la séquence aléatoire et la série de temps idéale est calculée. La plus grande statistique par itération produit un point dans la distribution empirique. La probabilité que la valeur de la statistique observée soit plus petite ou égale à une valeur quelconque est égale à l'étendue de la valeur dans la distribution nulle divisée par le nombre de points dans la distribution. Celle-là est la probabilité que la statistique maximale produite dans une expérience où le stress physiologique n'a aucun effet soit plus petite ou égale que la corrélation observée (Belmonte & Yurgelun-Todd, 2001).

Le test se réalise sur 3 phases tel que montré dans la Figure 6.4. Dans la première phase, la statistique du test (soit les coefficients de corrélation ou la valeur t de la statistique du test de différence des moyennes) est calculée. Dans la deuxième phase il s'agit de construire la distribution nulle en choisissant pour N itérations un arrangement aléatoire des valeurs t sur lesquels la statistique originale a été calculée. C'est la séquence du temps qui est randomisée à chaque étape de cette phase. Pour tenir compte des tests multiples, seule la plus grande valeur

de la statistique durant chaque itération est retenue. La méthode est avantageuse car nous contrôlons à la fois deux ensembles de valeurs de signification: les estimations ponctuelles pour chaque SNP et une valeur qui contrôle le FWER, étant donné que chaque nouvelle valeur p (appelé valeur p ajustée) reflète la chance de trouver une statistique aussi grande étant donné que l'on a effectué le même nombre de tests que des SNPs et de phénotypes qu'on a.

Supposons qu'un sous-ensemble des SNPs U $\subseteq$ V, où V est l'ensemble total de SNPs. Pour tous les i, 0<=i<N, l'i^{e} élément dans la distribution des corrélations maximales sur cette région (R_U) doit être inférieure ou égale en magnitude à l'i^{e} élément de la distribution maximale des corrélations sur l'ensemble de l'image V (R_V). Ainsi, le rangement de toute corrélation r contre la distribution (R_U) doit toujours être au moins aussi loin dans les extrémités de la distribution que le rangement de toute corrélation r contre la distribution (R_V). Les niveaux de signification calculés pour chaque marqueur dans V par rapport à l'ensemble V au complet ne pourraient jamais dépasser celles des marqueurs dans l'ensemble partielle U. Par conséquent, le test de permutation contrôle fortement le FWER.

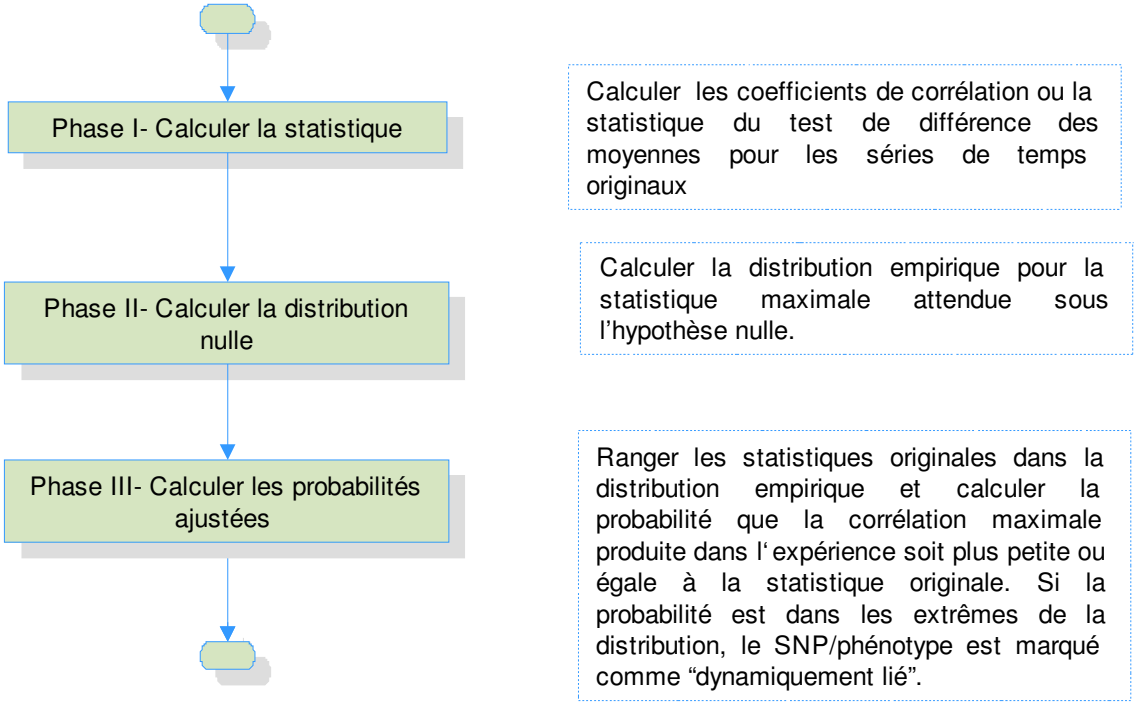


Figure 6.4 Représentation des tests de permutation pour l'analyse de liaison dynamique

Dans la troisième phase, les statistiques originales sont comparées avec la nouvelle distribution nulle. Une nouvelle valeur p est calculée qui reflète la probabilité que la statistique maximale

produite dans l'expérience soit plus petite ou égale à la statistique originale. Selon l'expérience de Belmonte, (Belmonte & Yurgelun-Todd, 2001) , le fait de permuter toutes les séries de temps en réarrangeant les temps permet de conserver les effets des corrélations entre les mesures répétées et de la corrélation entre séries de valeurs t appartenant à SNPs proches ou très liées.

Le test de différences des moyennes appariées entre deux échantillons a été introduit pour tester l'hypothèse nulle que la moyenne des différences des statistiques du test de liaison (les valeurs t) entre les deux périodes est significativement égale à zéro. Apparier les mesures implique que l'on veut minimiser sa dissimilarité sauf pour le facteur en question, soit l'effet des tests physiologiques. Dans notre cas, les mesures sont appariées pour minimiser la variation entre les mesures dans chaque position en maximisant la variation des mesures d'une position à l'autre. De cette manière, la différence entre les réponses aux stimuli est attribuée au changement de position et non à la différence entre les mesures individuelles durant le test. Étant donné que le nombre de mesures appariées est très petit, des violations aux présuppositions de données aberrantes et des éloignements de la distribution gaussienne sont surement présentes. La méthode de permutation nous permet donc de calculer la distribution des mesures sous l'hypothèse nulle et de quantifier si l'effet est vérifiable en permutant les mesures dans temps ou non. Le test de différences appariées est recommandé parce que les mesures répétées sont fortement corrélées et l'on peut gagner en précision même si l'on perd en degrés de liberté.

6.4.2 Simulations pour l'analyse de liaison dynamique

Lors de l'interprétation de résultats des analyses sur les généalogies, il est extrêmement utile de savoir combien de fois un résultat similaire pourrait se produire par hasard. Par exemple, dans notre analyse de liaison il est utile de savoir combien des résultats significatifs sont attendus par rapport à l'ensemble des phénotypes et aux marqueurs génétiques disponibles. Si l'on a des doutes sur la liaison de certains génotypes suspects, il est important de caractériser le faux-positifs et d'établir le taux d'erreur de détection. Au centre de Recherche du CHUM un module a été implanté et est utilisé pour simuler les génotypes dans les tests de permutation en se servant du logiciel Merlin (Abecasis, Cherny, Cookson, & Cardon, 2002). Merlin effectue ses simulations en remplaçant les données génotypiques d'entrée avec des nouveaux marqueurs

conditionnels à la structure familiale, à la fréquence des génotypes et à la distribution des données manquantes.

Merlin utilise la méthode de *gene-dropping* en simulant la transmission des allèles des parents aux enfants. La méthode suppose qu'aucune force extérieure telle que la sélection, les mutations ou la pression sélective agissent au locus en question, donc le locus est en équilibre de Hardy Weinberg. Ainsi, la transmission des allèles de chaque parent au même enfant est indépendante ainsi que les transmissions des parents aux autres enfants. Les allèles sont donc transmis selon les lois mendéliennes. Selon (Brunelle, 2008) une bonne méthode de rééchantillonnage doit respecter le plus possible les caractéristiques des données originales. La méthode de rééchantillonnage implantée respecte la transmission mendélienne des génotypes, la distribution des valeurs manquantes et la fréquence observée des allèles dans la population.

Merlin (Abecasis, et al., 2002) procède en trois phases :

1- Assigne des chromosomes aléatoires aux parents en respectant les fréquences alléliques de chaque marqueur. Par défaut les allèles sont simulés indépendamment, mais le logiciel a l'option de tenir compte du déséquilibre de liaison.
2- Les allèles des enfants sont transmis en tenant compte des relations dans les données originales et de la fraction de recombinaison observée.
3- Les génotypes originaux sont remplacés par les simulés en retenant le même pattern de données manquantes. Par exemple, si le dans le marqueur B le génotype de l'individu i est manquant, il le sera dans tous les échantillons produits à partir de cet ensemble.

La valeur p de signification du test de permutation par chaque période est déterminée en comptant le nombre de fois que le minimum des valeurs t des tests de liaison de la période obtenu par simulations est plus grand ou égal que le minimum des valeurs t des tests de liaison pour la période sur les données originales. Ce nombre divisé par le total de simulations effectuées devient la valeur p du test de liaison par période. Le choix du minimum des valeurs t de la statistique par période nous permet d'avoir un contrôle fort du FWER: aucun point dans un sous-ensemble des simulations par période ne sera plus petit que celui choisi pour calculer la signification du test.

La valeur p du test de liaison dynamique est obtenue en comptant combien de fois la magnitude d'une statistique donnée est plus grande ou égale dans las simulations que dans l'expérience

originale. Ce nombre divisé par le totale de simulations effectués devient la valeur p du test de liaison dynamique. Nous retenons deux statistiques: la différence des minimums et la différence de la statistique t du test de différences des moyennes. Le test des différences des minimums calcule le nombre de fois que la différence entre les minimums des valeurs t obtenus par simulation dans chaque période est plus grande en magnitude que la différence entre les minimums observés dans chaque période. Le test de différence des moyennes calcule le nombre de fois que la magnitude de la valeur t d'un test de différences des moyennes entre les valeurs t de la position couché et les valeurs t de la position debout est plus grande ou égale dans les simulations que la même statistique calculé sur les données originales.

Les simulations sont très coûteuses du point de vue du temps de calcul. En général, on ne teste que les paires marqueur-phénotype plus significatives et le nombre de simulations à effectuer est calculé vis-à-vis les valeurs p observées. Sur ces ensembles choisis, on devrait confirmer les résultats des tests de liaison, en plus de déterminer la valeur p empirique durant une position orthostatique donnée.

6.5 Signification des tests d'association génétique

La réduction de dimensionnalité des phénotypes est proposée dans le but de diminuer le nombre de tests effectués en se servant de méthodes qui permettent d'obtenir une meilleure représentation des phénotypes vis-à-vis les tests physiologiques, à la variabilité due aux mesures répétées et à la variabilité due à la corrélation entre les phénotypes. La Figure 6.5 et la Figure 6.6 montrent la matrice corrélation entre les variables dans les positions couché et debout, respectivement. Plus la magnitude de la corrélation est haute, plus la couleur est intense. Outre les statistiques descriptives qui nous permettent d'obtenir une représentation généraliste des donnés, deux autres méthodes ont été choisies parce qu'ils tiennent compte de la corrélation entre les phénotypes et modélisent la variabilité: l'analyse de Composantes principales et l'analyse factorielle. Nous nous attendons à ce que les mesures répétées des multiples phénotypes soumises à des méthodes de réduction produisent quelques variables indépendantes qui représentent en bonne mesure la variabilité totale et la corrélation entre les variables originales.

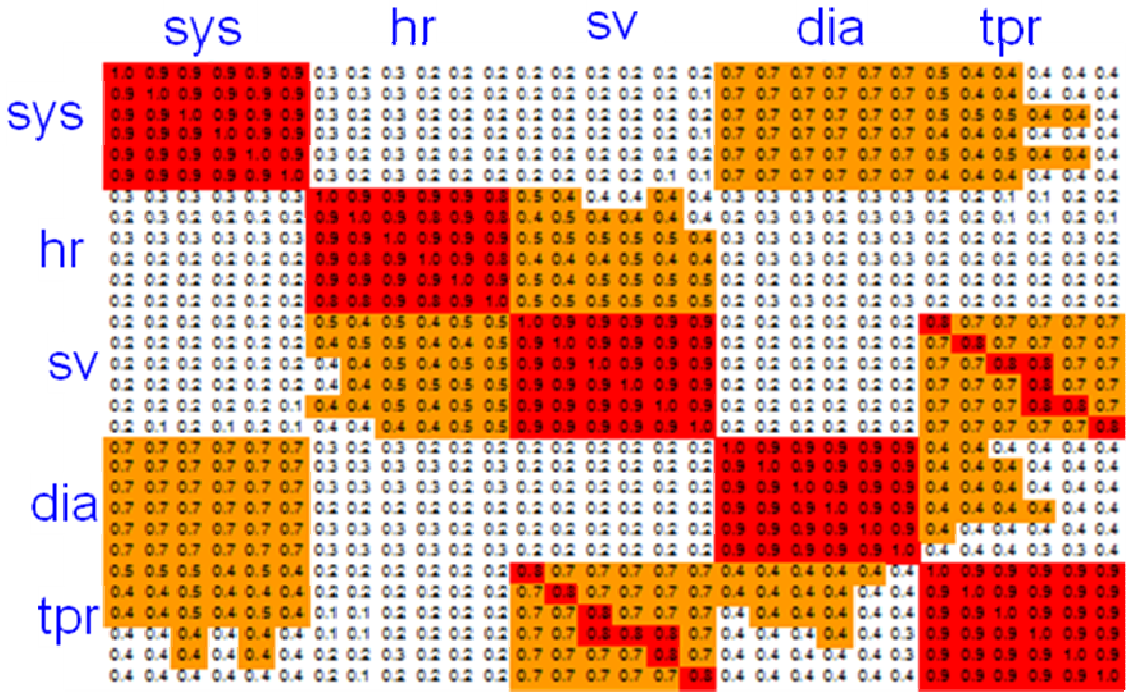


Figure 6.5 Matrice de corrélation des traits intermédiaires de la pression sanguine pendant la position couchée (temps 30 à 55) du jour 1.

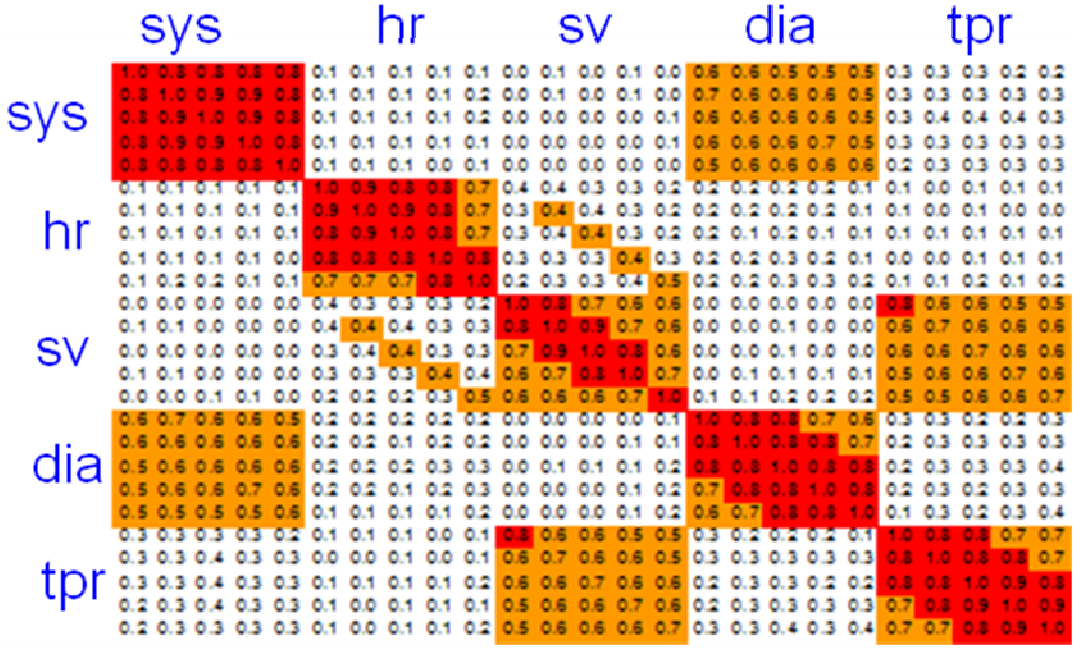


Figure 6.6 Matrice de corrélation des traits intermédiaires de la pression sanguine pendant la position debout (temps 62 à 70) du jour 1.

En conséquence les interactions entre les variables sont modélisées et nous obtenons un ensemble optimisé de variables indépendantes qui encapsulent la variance totale des variables originales, en se basant principalement sur les mesures de covariance ou corrélation entre les phénotypes en question.

6.5.1 Simulations pour la comparaison des méthodes de réduction de dimensionnalité

Une simulation a été implantée dans le but de comparer les méthodes de réduction de dimensionnalité. Étant donné que l'on ne vise pas d'ajuster les valeurs p de l'association, mais plutôt de comparer l'efficacité des méthodes de réduction, nous avons décidé de simuler les phénotypes. Une telle méthode est très libérale pour être utilisé pour corriger les valeurs p des tests d'association étant donné qu'elle ne tient pas compte de la corrélation phénotypique entre les phénotypes des individus généalogiquement rapprochés. Une permutation libre des phénotypes implique que les sujets et les génotypes et phénotypes sont indépendants les uns des autres. Dans le contexte des tests d'association familiale, une telle supposition implique que le test de l'hypothèse qui corrige pour les liens au sein d'une fratrie devient conservateur.

Selon Balding (Balding, 2006) lorsque les tests sont faiblement significatifs on peut se servir des permutations pour estimer l'erreur de type I pour son ensemble de données particulier. On pourrait en conséquence relaxer ses seuils de signification. Tel que montré dans la Figure 6.7, à chaque itération des phénotypes aléatoires sont réassignés aux individus et des phénotypes dérivés sont recalculés ainsi que les tests d'association familiale. Ces nouveaux phénotypes sont distribués selon une loi gaussienne multidimensionnelle de moyenne égale à la moyenne observée des données et respectent la structure de corrélation des phénotypes observés (ils sont construits à partir de la matrice de covariances des données originales). Deux types de seuils peuvent être estimés à partir de ces résultats: le premier est un seuil Z_0 qui peut être estimé séparément pour chaque point dans l'expérience et qui fournit la valeur critique à 100(1 - α) % à chaque test. Le deuxième type de seuil est un seuil au niveau de toute l'expérience, qui fournit une valeur critique à 100(1 - α) %. Cette valeur critique est valide simultanément pour tous les points dans l'analyse. La valeur critique Z_0 est égale au percentile 95% sur la distribution des valeurs Z maximales (Churchill & Doerge, 1994). Cette valeur critique nous permet de détecter une association quelque part dans l'expérience tout en contrôlant que le taux d'erreurs de type I soit plus petite ou égal que α= 0.05. Le seuil de signification ajusté pour

l'ensemble de données est approximée par $\alpha_{adj}=1 - 2*(1-p_N(Z_0))$, étant $p_N(Z_0)$ la fonction de répartition pour la valeur critique Z_0 en sachant que la statistique du test FBAT suit approximativement une loi gaussienne de moyenne égale a 0 et variance égale à 1 et que pour le test d'association l'hypothèse alternative est bilatéral.

Le taux d'erreur de type I est calculé, défini comme le nombre d'essais donnant une valeur p plus petite que le seuil établi (0.05). À la fin la capacité de chaque méthode pour détecter des fausses associations sera établie et les méthodes seront comparées par rapport à cette mesure.

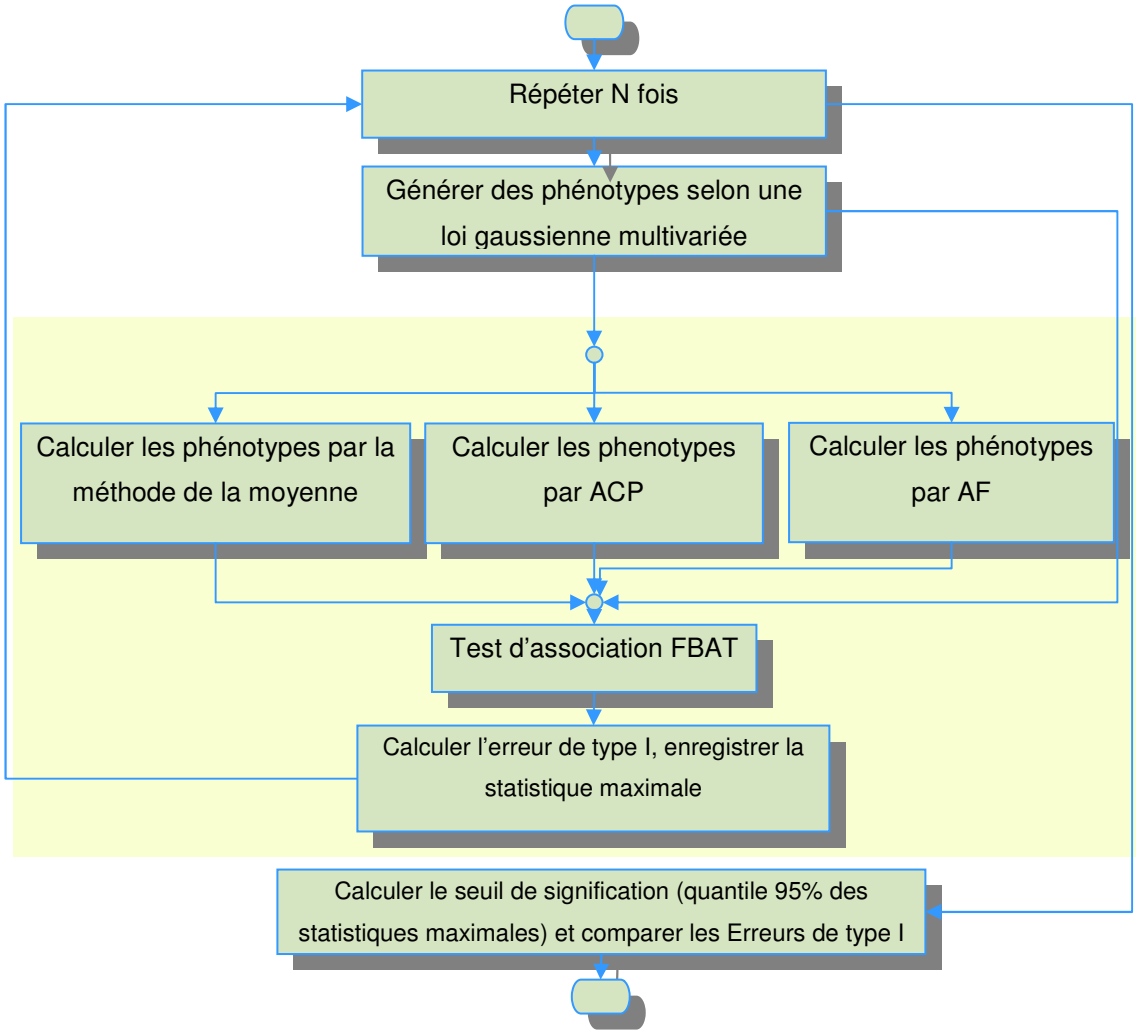


Figure 6.7 Diagramme de flux pour les tests de permutation quantifiant les différences entre les méthodes de réduction de dimensionnalité

CHAPITRE 7. ASPECTS INFORMATIQUES

Au sein de l'équipe du Centre de recherche du CHUM où ce projet à été effectué, on offre aux chercheurs une absolue liberté pour le choix de plateformes et des logiciels à utiliser, ainsi que pour les méthodologies de conception. Il n'existe pas une application principale sur laquelle les bioinformaticiens contribuent, mais une série de scripts génériques en Python servant à accéder à la base de données CARDIOGEN, fournir des structures de données standard et exécuter plusieurs analyses allant des tests statistiques standard jusqu'aux tests de permutation poussés. L'application CARDIOGEN (du même nom que la base de données) est principalement utilisée pour l'extraction de l'information génotypique et phénotypique des Canadiens Français et très rarement pour les analyses statistiques sur ces données. Les logiciels construits s'inspirent donc de la fonctionnalité existante et ont été conçus par rapport à ces programmes-là.

Dans ce chapitre nous présentons les principales activités informatiques que nous avons menées: la conception du logiciel, son implantation et sa validation. Le développement du logiciel a été principalement guidé par deux critères: le développement rapide et la vitesse d'exécution. Le développement rapide permettra de réaliser toutes les analyses demandés en permettant de changer rapidement de voie si la méthode proposée n'est pas valide ou est peu puissante (Brunelle, 2008). L'exécution du logiciel doit être assez rapide pour produire des résultats dans un délai raisonnable.

Toutes les analyses ont été implantées en R et Python. Python est un langage de choix pour le développement rapide. Il s'apprend rapidement et le code écrit en Python est compact et facile à comprendre. Ainsi, il possède un niveau d'abstraction élevée et une librairie riche. Python fournit aussi des structures de donnés simples tels que les listes et les tables de hachage en permettant d'utiliser la même syntaxe pour itérer sur une liste ou une table de hachage. R est un progiciel d'analyse statistique et graphique, disponible gratuitement pour les plateformes Windows, Mac OS X et Linux. R est à la fois le langage de programmation et le progiciel de fonctions statistiques. Des nombreux contributeurs de tous les domaines ajoutent des fonctions des plus simples aux plus complexes et ses librairies standard fournissent une grande partie des tests statistiques utilisés dans cette étude. L'intégration des logiciels construits en R est garantie par la librairie rpy de Python. Rpy est l'interface entre R et Python et permet de gérer les différents types d'objets R, ainsi que d'exécuter autant les fonctions standard de R que n'importe

quelle fonction arbitraire. Rpy assure aussi le formatage des structures de données Python en structures de données R et vice-versa.

Les sections sont divisées selon les expériences réalisées, et à l'intérieur de chaque expérience les sections sont organisées selon l'activité. La conception des analyses est expliquée par des vues de décomposition du système et dans certains cas par des diagrammes de flux. Les aspects d'implantation expliquent très sommairement le choix des langages et la section de validation explique la manière dont la méthode a été validée et comment nous nous sommes assurés que les méthodes implantés sont fiables.

7.1 Tests de permutation pour la liaison dynamique

Le développement rapide est une condition très importante à laquelle la méthode de permutation doit s'ajuster. Ainsi, si l'effet dynamique n'est pas présent dans l'ensemble de tests de liaison, on ne fera pas de différence dans les méthodes par rapport aux tests physiologiques et les tests de signification procéderont plutôt sur le total des mesures répétées. Les sections suivantes décrivent les activités de génie logiciel qui ont été réalisées.

7.1.1 Définition de requis

Nous avons produit un document de spécification de besoins accordé au standard IEEE, dont le document est inclus en annexe (Voir ANNEXE 2).

7.1.2 Conception

La Figure 7.1 montre la vue d'ensemble du test de permutation pour l'analyse de liaison dynamique. Une personne non expérimentée pourrait exécuter le test de permutation à condition que le format et le contenu des fichiers soient respectés. Le test se réalise en 3 phases et requiert les fonctionnalités suivantes:

- Un arbre de recherche pour enregistrer la plus grande statistique de façon ordonnée, ainsi que le SNP qui a généré cette statistique. L'arbre binaire AVL (arbre binaire de recherche équilibré) a été privilégié. Dans un arbre AVL, les hauteurs des deux sous-arbres d'un même nœud diffèrent au plus d'un. La recherche, l'insertion et la suppression sont toutes en *O (ln n)* dans le pire des cas. Les classes AVLTree et Tree de la librairie Opus7 fournissent la fonctionnalité requise (Preiss, 2005).

- Un module statistique pour le calcul des coefficients de corrélation et des statistiques de différence de moyennes. Le premier choix avant l'implantation et les tests était R, mais des problèmes d'efficacité de l'interface entre Python et R (rpy) ont forcé l'utilisation de la librairie scipy de Python (Jones, 2001) .

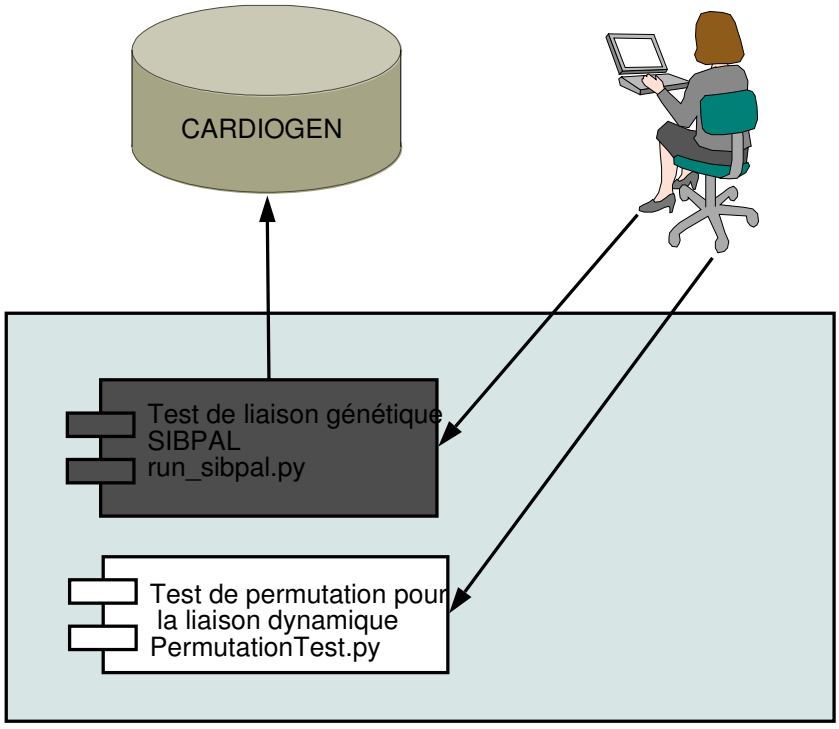


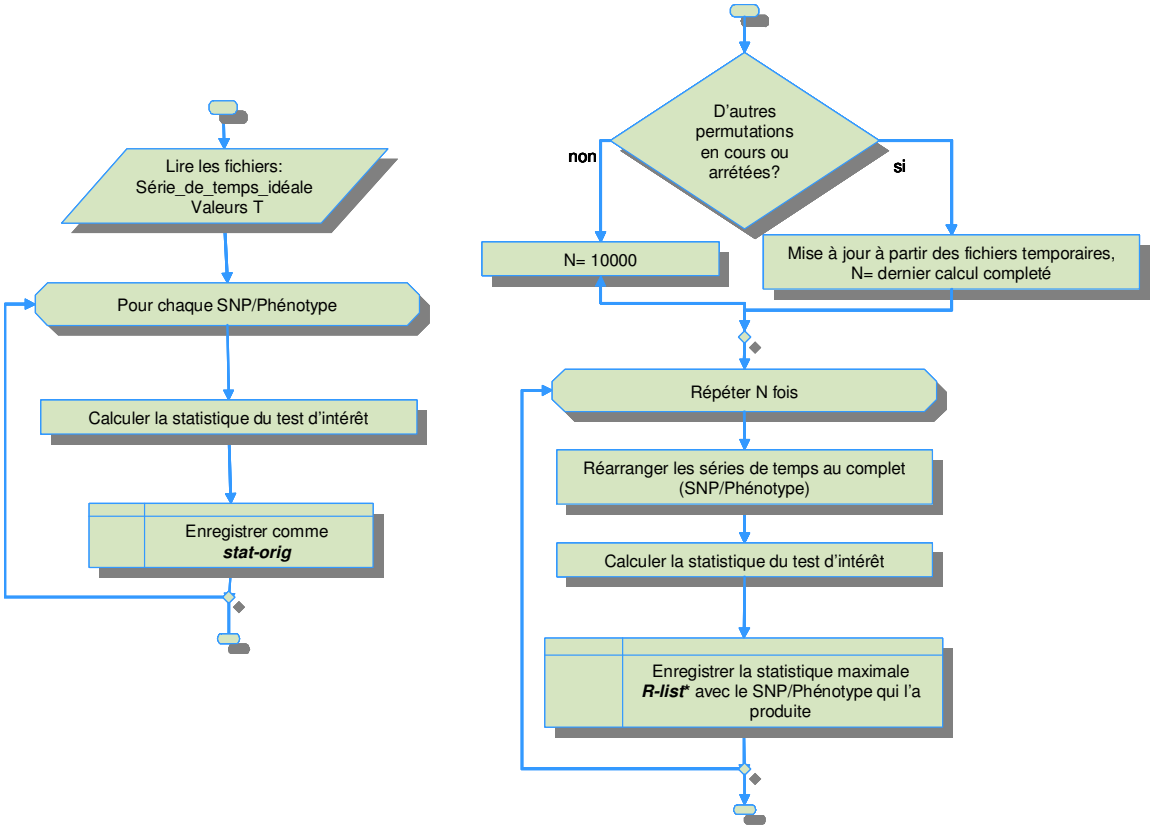
Figure 7.1 Vue de décomposition pour le test de permutation pour l'analyse de liaison dynamique. La couleur la plus foncée indique que le module existe. La couleur plus claire indique qu'il s'agit d'un nouveau module.

La Figure 7.2 (a) montre le diagramme de flux pour la phase I du test de permutation. D'abord, la statistique d'intérêt est calculée sur les données originales (des séries de temps de valeurs t

issues du test de liaison point par point). Les valeurs de la statistique sont enregistrées dans un arbre binaire dont la clé est la valeur de la statistique calculée (*stat-orig*). Le nœud contient aussi le SNP, phénotype et la valeur p du test effectué).

Dans la phase II on procède par une méthode itérative. À chaque itération:

- les temps de mesure sont réarrangés aléatoirement,
- les statistiques du test d'intérêt sont calculées et la plus grande valeur de la statistique est enregistrée dans un arbre binaire R.



(a) (b)

Figure 7.2 Diagramme de flux pour les phases I(a) et II(b) du test de permutation pour la liaison dynamique

Selon Belmonte (Belmonte & Yurgelun-Todd, 2001), le fait de découvrir qu'une statistique est significative après permutation pose un problème, étant donné la structure complexe de corrélation qui peut exister entre les séries de temps appartenant à SNPs proches ou en déséquilibre de liaison. La Phase II inclut alors la définition d'une distribution-substitut qui fera

que le test devient plus sensible à la présence de signaux corrélés. Une fois que le SNP a été déclaré comme dynamiquement liée, la valeur ou les valeurs de la statistique maximale que ce SNP a apporté à la distribution nulle sont remplacées par la valeur ou les valeurs plus grandes présentes dans la distribution substitut. La distribution nulle peut donc varier dans la phase III.
Le nombre total de valeurs apportées à chaque itération sur la distribution substitut est selon leur expérience de deux, mais le programme devrait permettre la paramétrisation de cette valeur.

Également, tel qu'annoncé dans la Figure 7.2(b), un mécanisme de réponse suite à une défaillance ou à une terminaison anormale de l'application a été établi. Il s'agit d'enregistrer dans un fichier les valeurs contenues dans l'arbre la distribution nulle et dans celui de la distribution substitut. Au fur et à mesure que les statistiques sont calculées, deux fichiers de sortie sont créés et mis à jour. La Phase II vérifie si dans le répertoire courant de l'exécution une analyse a pris fin abruptement. Si c'est le cas, l'information déjà calculée est chargée dans les structures définies et la phase II recommence à partir de la dernière qui a été complétée (le nombre de lignes bien formatées du fichier temporaire correspondant à la distribution nulle). Ces fichiers temporaires sont effacés à la fin de la phase III.

Dans la phase III, représenté dans la Figure 7.3, la liste de probabilités ajustées pour chaque SNP est d'abord initialisée avec la valeur 0.5 pour tous les SNPs. Ensuite, la liste de statistiques produite dans la Phase I est ordonnée. On extrait itérativement la statistique maximale avec le SNP qui l'a produite. La corrélation est classée dans la distribution empirique calculée dans la phase II. Ce classement est projeté sur l'intervalle [0,1] afin de trouver la probabilité que la corrélation maximale produite, dans un espace de SNPs non liés dynamiquement, soit plus petite ou égale à celle produite pour ce SNP-ci. Si la probabilité ajustée est plus petite ou égale que le seuil α (par défaut 0.05, pouvant varier selon le choix de l'utilisateur) divisé par deux ou $1-\alpha/2$, alors le SNP est marqué comme lié dynamiquement. D'un autre côté, les valeurs de la corrélation maximale pour ce dernier sont effacées de la distribution nulle construite dans la phase II et sont remplacées pour les corrélations substitutes. Ce procédé-ci est répété jusqu'à ce que la probabilité ajustée de la statistique maximale extraite devienne non significative.

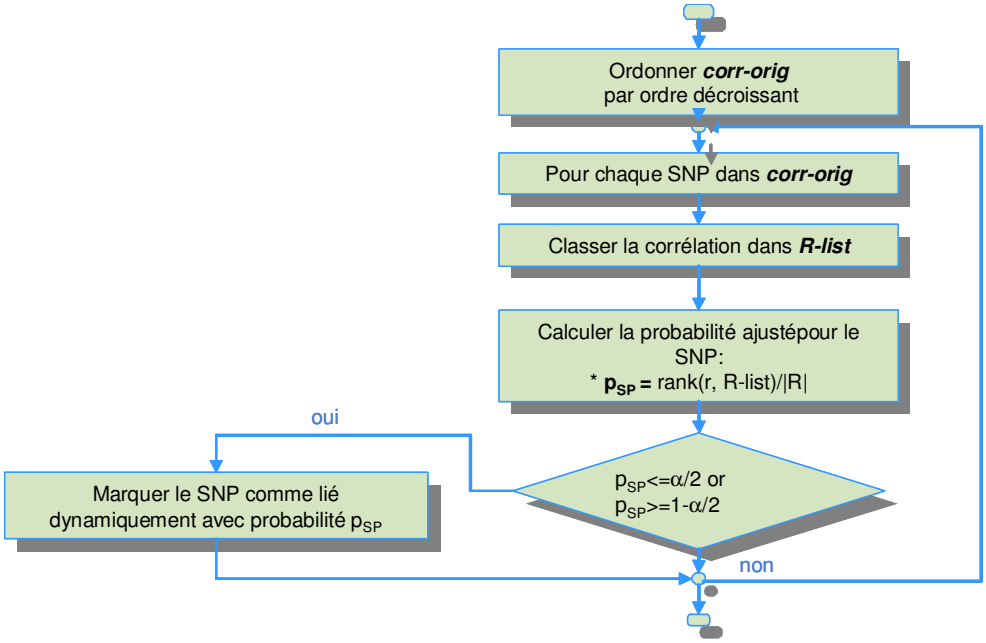


Figure 7.3 Diagramme de flux pour la phase III du test de permutation pour la liaison dynamique

Enfin, un fichier de sortie doit être créé contenant la liste de SNPs, la valeur de la statistique originale, la valeur p du test ajustée, la probabilité calculée dans la phase III et l'indicateur de liaison dynamique (oui ou non).

Une seule classe appelée PermutationTest contient la fonctionnalité du test de permutation, tel que montré dans la Figure 7.4. La Classe AVLTree fournit la fonctionnalité des arbres binaires. Trois structures du type AVLTree contiennent les valeurs des statistiques calculées dans les phases I et II (ActualCor, NulDistr et NulDistrSus). Des tables de hachage permettent d'accéder à l'information contenue dans les fichiers d'entrée (IdealTs, TVals, SNPList, dtList). D'autres structures sont nécessaires pour contenir l'information des SNPs qui ont produit les plus grandes valeurs de la statistique et permettre d'effacer des nœuds des arbres binaires NulDistr et NulDistrSus (InverseCorList et NullDistrTies). Finalement, la table de hachage DynLinkList contient l'information de liaison dynamique calculée dans la phase III. Les fonctions compute_actual_cor, compute_null_distr et compute_dynamic_linkage exécutent les phases I, II et II respectivement. Les fonctions read_tvalues et read_Idealts transfèrent l'information des fichiers à la structure de la classe. Les fonctions get_idealts, get_actual_correlations et get_null_distribution permettent d'accéder aux attributs de la classe dans un ordre spécifique à l'application (ordre inverse de la statistique pour les statistiques observées (corrélations courantes) et en ordre croissant pour la distribution nulle. La fonction

withdraw_from_null_distribution permet d'éliminer des nœuds de l'arbre des statistiques de la distribution nulle. Finalement insert_into_tree encapsule l'accès aux arbres binaires et enregistre au besoin la statistique obtenue dans le dictionnaire de statistiques par SNP (InverseCorList).

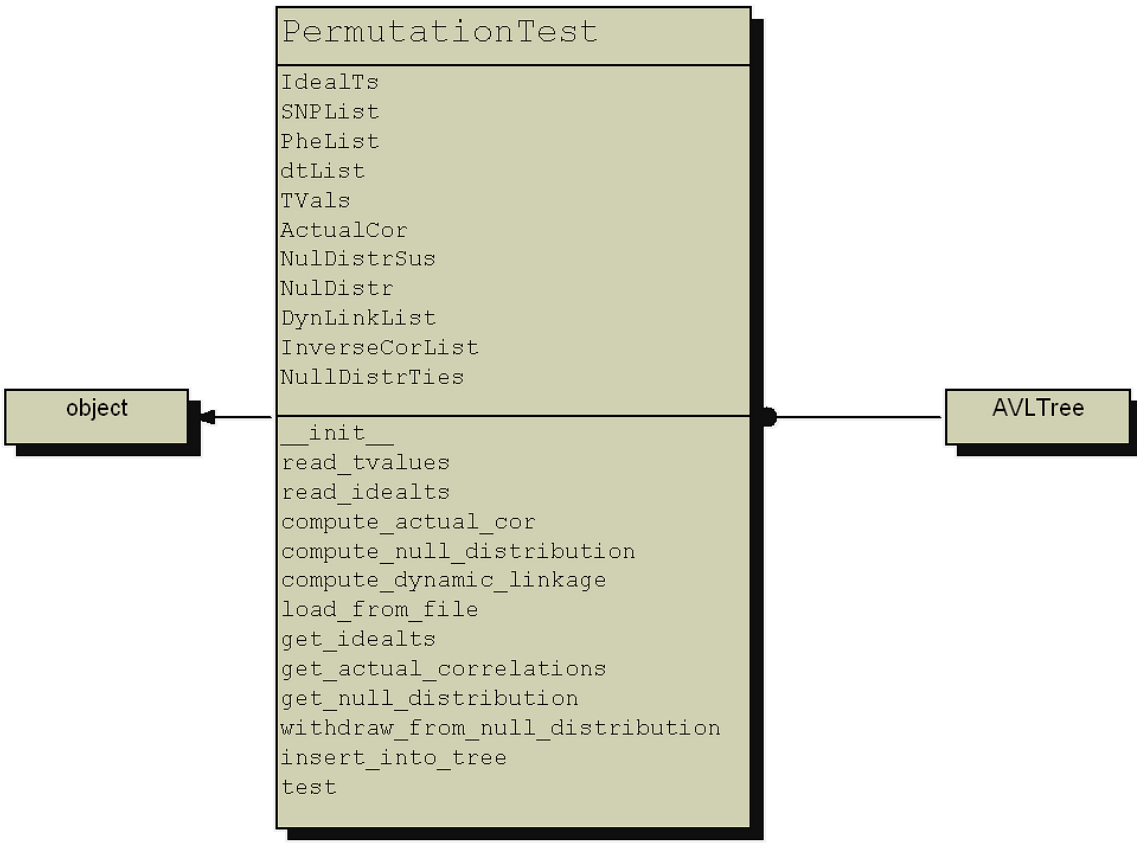


Figure 7.4 Diagramme de classes pour le test de permutation pour la liaison dynamique.

7.1.3 Aspects de l'implantation

Le test de Permutation à été implanté pour rouler dans plateformes Windows et Linux, les fichiers d'entrée doivent respecter une certaine logique et l'application ne dépend pas des librairies et modules d'accès restreints. Pour les arbres binaires requis nous avons utilisé les classes AVLTree et Tree de la librairie Opus7 (Preiss, 2005). Le calcul des statistiques pour le même test (les coefficients de corrélation et des statistiques de différence de moyennes) a été implanté initialement en R. Dans le but de diminuer le temps d'exécution du test de permutation nous avons utilisé la librairie *scipy* de Python (Jones, 2001), ce qui a diminué le temps

d'exécution de 9 heures à 42 minutes pour 1000 SNPs testés sur un phénotype dans un ordinateur PC.

7.1.4 Complexité algorithmique

Soit T le nombre de mesures de chaque série de temps, N le nombre de marqueurs et M la taille de la distribution empirique (le nombre de permutations). Les tests de corrélation et de différences des moyennes sont linéaires en T et l'insertion des statistiques dans l'arbre binaire se fait en O(log(N)). Chacune des opérations précédentes est réalisée une fois pour chaque marqueur, donc la phase I se réalise dans O(NT + Nlog(N)).
Dans la phase II les permutations de la série de temps sont linéaires en T (nous permutons les étiquettes du temps) et la série de temps est accédée en O(1). Le calcul des corrélations ou des tests de différences des moyennes se fait en O(NT). Enregistrer la statistique maximale dans l'arbre binaire prend O(log(M)). Chacune des opérations précédentes est réalisée M fois. La phase II prend O(MNT+Mlog(M)). Dans la phase III l'extraction de la statistique maximale de l'arbre binaire des statistiques observées prend O(log(N)). La localisation de cette statistique dans la distribution empirique prend log(M). L'élimination de la distribution nulle des statistiques produites par le marqueur en question prend O(log(M)) (on utilise un tableau de hachage pour récupérer les statistiques maximales obtenues par le marqueur en question), donc les opérations de recherche et élimination se font toujours en O(log(M)). Chacune des dernières opérations se réalise une fois pour chaque marqueur dont le test sur la statistique observée est significatif, ce qui est en une fraction de N. La Phase III est réalisée en O(Nlog(N) + Nlog(M)). Le test de permutation au complet est réalisé en O(MNT+M log(M)s + Nlog(N)).

Selon Belmonte (Belmonte & Yurgelun-Todd, 2001) étant le test linéaire en MNT, il est recommandable de restreindre M,N, et T aux valeurs minimales nécessaires pour produire des statistiques adéquates. Il recommande de ne pas faire moins de 10000 itérations pour éviter que la distribution empirique soit trop granulaire. Le seul choix c'est de diminuer N sans risque de perte d'information en réduisant l'ensemble de marqueurs selon un critère spécifique. Nous avons certainement réduit l'ensemble en ne choisissant que des marqueurs qui sont sur des gènes et qui sont candidats à être liés aux phénotypes intermédiaires de l'hypertension. Une autre réduction possible pourrait être l'élimination de marqueurs en déséquilibre de liaison avec d'autres, mais le test est robuste par rapport à la corrélation entre les séries de temps et nous risquons de perdre de l'information utile.

7.2 Simulations pour l'analyse de liaison dynamique

Au centre de Recherche du CHUM un module est utilisé pour simuler les génotypes dans les tests de permutation en se servant du logiciel Merlin (Abecasis, Cherny, Cookson, & Cardon, 2002). La méthode de rééchantillonnage implantée respecte la transmission mendélienne des génotypes, la distribution des valeurs manquantes et la fréquence observée des allèles dans la population. Un fichier avec la liste de paires marqueur-phénotype que l'on veut tester doit être préparé avant l'exécution des simulations. Une fois les tests de permutation on été exécutés, un programme extrait les résultats des tests fournis par SIBPAL (les fichiers en format .treg) et créée un fichier contenant dans chaque ligne les résultats du test de permutation pour un temps donnée. Deus fichiers sont crées: le fichier avec extension .pvals contient les valeurs p des tests sur les génotypes simulés. Le fichier avec extension .actuals contient les résultats du test de liaison sur les données originales.

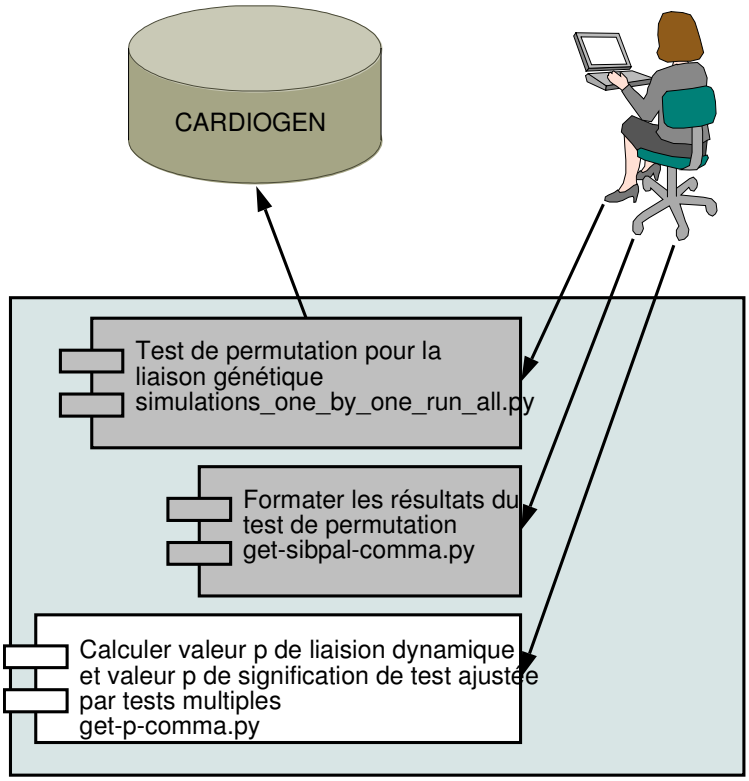


Figure 7.5 Vue de décomposition pour les simulations pour l'analyse de liaison dynamique. La couleur la plus foncée indique que le module existe. La couleur plus claire indique qu'il s'agit d'un nouveau module, la couleur intermédiaire indique une modification de fonctionnalité existante.

Le module de simulations_one_by_one a été modifié pour inclure des mesures répétées des phénotypes dans la procédure de sélection des phénotypes à extraire et les covariables ont été associés à chaque liste de mesures répétées.

Le module qui exécute le formatage des résultats des tests de permutation a été modifié pour assurer que lorsque le test de liaison a un résultat no valide une valeur par défaut est associée au résultat du test.

Le module de calcul de la valeur p de liaison dynamique et la valeur p de signification des tests ajustés par tests multiples réalise les tâches suivantes:

- Parcours le dossier d'entrée et extrait des paires de fichiers .pvals et .actuals les valeurs t des tests de liaison pour les données originales et pour les permutations (la statistique t du test de liaison).
- Obtient la valeur p du test de liaison dynamique en comptant combien de fois la magnitude d'une statistique donnée est plus grande ou égale dans las simulations que dans l'expérience originale. Ce nombre divisé par le totale de simulations effectués devient la valeur p du test de liaison dynamique. Plusieurs tests ont été désignés, mais seulement les résultats des tests MIN et différence des moyennes ont été retenus. Le test MIN calcule le nombre de fois que la différence entre les minimums des valeurs t obtenus par simulation dans chaque période est plus grande en magnitude que la différence entre les minimums observés dans chaque période. Le test de différence des moyennes calcule le nombre de fois que la magnitude de la valeur t d'un test de différences des moyennes entre les valeurs t de la position couché et les valeurs t de la position debout est plus grande ou égale dans les simulations que la même statistique calculé sur les données originales.
- La valeur p de signification du test de permutation par chaque période est déterminée en comptant le nombre de fois que le minimum des valeurs T des tests de liaison de la période obtenu par simulations est plus grand ou égal que le minimum des valeurs t des tests de liaison pour la période sur les données originales. Ce nombre divisé par le total de simulations effectuées devient la valeur p du test de liaison par période. Pour le calcul de la statistique du test de moyennes indépendantes nous utilisons le module *stats* du package la librairie Scipy de Python (Jones, 2001) .

7.3 Analyse d'association familiale avec réduction de dimensionnalité des phénotypes

Pour les méthodes de réduction de dimensionnalité, le logiciel R fournit les fonctions *factanal* et *princomp* pour l'analyse de composantes principales et l'analyse factorielle respectivement (R-Development-Core-Team, 2007). FBAT réalise les analyses d'association familiale et requiert d'un formatage spécifique des fichiers d'entrée. Le module pour l'exécution des tests d'association avec FBAT existe et une modification a été introduite pour permettre de lire les fichiers de pedigree et les génotypes à partir d'un fichier de texte plain. Cela à fin d'éviter la lecture répétée des génotypes dans la base de donnés, surtout au moment de la simulation. En conséquence, il s'agit d'implanter les modules de traitement des phénotypes et la comparaison des résultats des tests et d'adapter le module le module d'appel aux tests d'association avec FBAT. La
Figure 7.6 montre une vue d'ensemble du système.

Le module d'extraction de phénotypes est fourni pour l'application CARDIOGEN, les individus avec des phénotypes manquants doivent être élimines parce que l'analyse de composantes principales ne peut pas traiter les données manquantes. Si un seul des phénotypes est manquant, l'application ne pourra pas traiter le fichier au complet. La définition des paramètres pour l'exécution de FBAT est faite en Python et ne peut pas être réalisée par une personne non expérimentée. Une analyse sur un nouveau fichier de phénotypes implique que la personne connaît la syntaxe de Python.

Le module d'analyse d'association avec réduction de dimensionnalité réalise les tâches suivantes:

- Calcule les phénotypes dérives en appelant des fonctions écrites en R selon la position dans les fichiers des phénotypes dans les positions debout et couchée.
- Exécute l'analyse d'association familiale entre les phénotypes et les génotypes observés (point par point) avec FBAT en utilisant le module run_fbat.py
- Exécute l'analyse d'association familiale pour les phénotypes dérivés avec FBAT en utilisant le module run_fbat.py pour les phénotypes dérivés.
- Calcule la correction des valeurs p pour les associations point par point, a fin de n'avoir qu'une valeur p par position.

- Apparie les résultats des tests suivants (une valeur p par phénotype, par SNP et par position) :
    - point par point corrigés par la méthode de Bonferroni,
    - phénotypes moyenne
    - phénotypes ACP (l'assignation d'une paire phénotype/composante principale se fait selon les facteurs de pondération de la variable par rapport à la composante principale. Le plus grand facteur de pondération indique que la composante principale explique plus la variabilité d'un phénotype qu'un autre)
    - phénotypes AF (l'assignation d'une paire phénotype/facteur se fait selon les facteurs de pondération de la variable par rapport au facteur).
- Compare les résultats des tests (test t sur les différences entre les méthodes, test de Wilcoxon et proportion de tests significatifs.

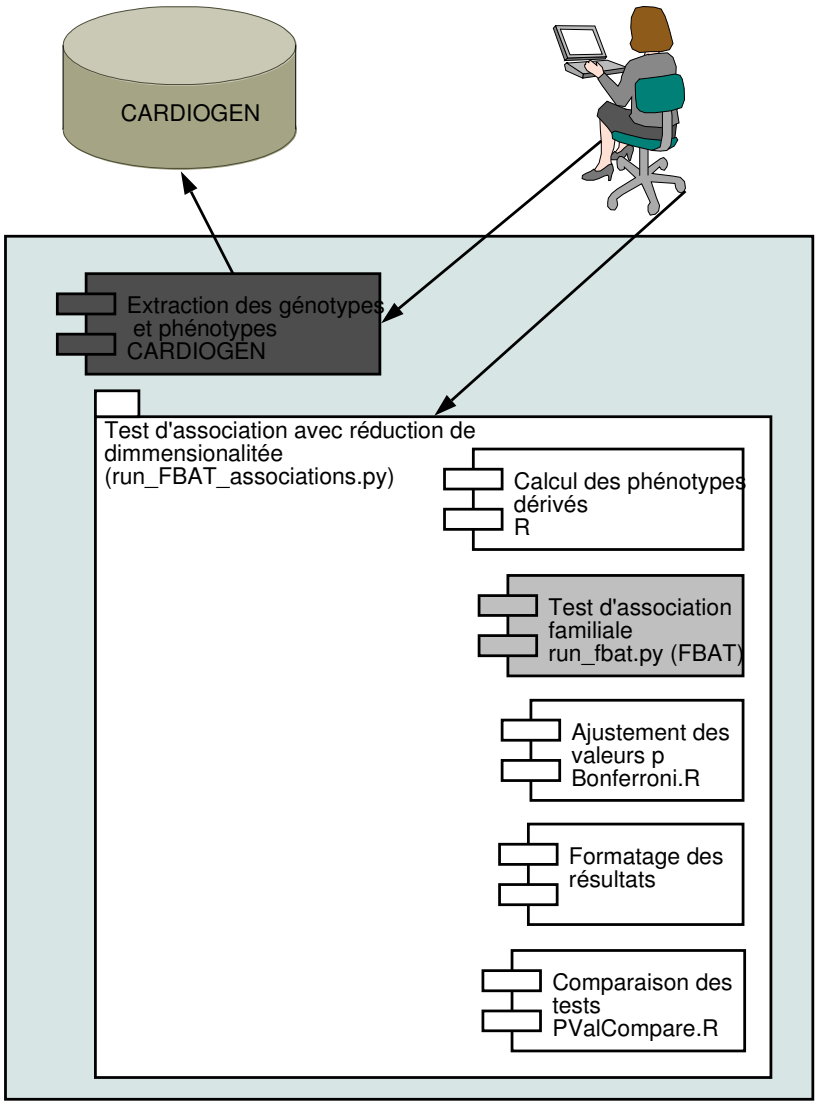


Figure 7.6 Vue de décomposition pour l'analyse d'association avec réduction de dimensionnalité des phénotypes pour l'analyse d'association familiale. La couleur la plus foncée indique que le module existe. La couleur plus claire indique qu'il s'agit d'un nouveau module, la couleur intermédiaire indique une modification de fonctionnalité existante.

7.4 Simulations pour la comparaison des méthodes de réduction de dimensionnalité

La Figure 7.7 montre une vue d'ensemble de la procédure de simulations pour la comparaison des méthodes de réduction de dimensionnalité et le calcul du seuil de signification empirique. Il s'agit d'adapter le module des tests d'association avec réduction de dimensionnalité des

phénotypes pour qu'il réalise les mêmes opérations sur le même ensemble de génotypes mais avec un nouveau set de phénotypes simulés. Tel que montré dans la Figure 6.7, à chaque itération des phénotypes aléatoires sont réassignés aux individus et des phénotypes dérivés sont recalculés ainsi que les tests d'association familiale. Ces nouveaux phénotypes sont distribués selon une loi gaussienne multidimensionnelle de moyennes égales à la moyenne observée des données et avec la matrice de covariance estimée à partir des phénotypes observés. Pour la génération des phénotypes simulés nous avons utilisé la fonction *rmvnorm* de la librairie *mvtnorm* de R (R-Development-Core-Team, 2007).

Selon Balding (Balding, 2006) lorsque les valeurs p des tests ne sont pas très significatives on peut se servir des permutations pour estimer l'erreur de type I pour un ensemble de données particulier. Le programme qui calcule le seuil de signification empirique pour les tests d'association avec réduction de dimensionnalité réalise les tâches suivantes:

- extrait les valeurs p des tests d'association avec des phénotypes simulés à partir d'un fichier contenant tous les résultats des tests d'association des phénotypes réduits (fbat-art.csv);
- se sert d'une table de hachage pour construire la distribution de la valeur t maximale obtenue pour chaque phénotype dans chaque simulation par période;
- écrit un fichier avec la liste de valeurs maximales du test t d'association par méthode par période.

À partir du fichier contenant la valeur maximale de la statistique t du test d'association par méthode par période on calcule le seuil de signification du test d'association pour chaque méthode par période. La valeur critique T_0 est égale au percentile 95% sur la distribution des valeurs t maximales. Cette valeur critique nous permet de détecter une association quelque part dans l'expérience tout en contrôlant que le taux d'erreurs de type I soit plus petite ou égal que $\alpha = 0.05$ (Churchill & Doerge, 1994). Le seuil de signification ajusté pour l'ensemble de données est approximée par $\alpha_{adj}=1 - 2*(1-p_N(T_0))$, étant $p_N(T_0)$ la fonction de répartition pour la valeur critique Z_0.

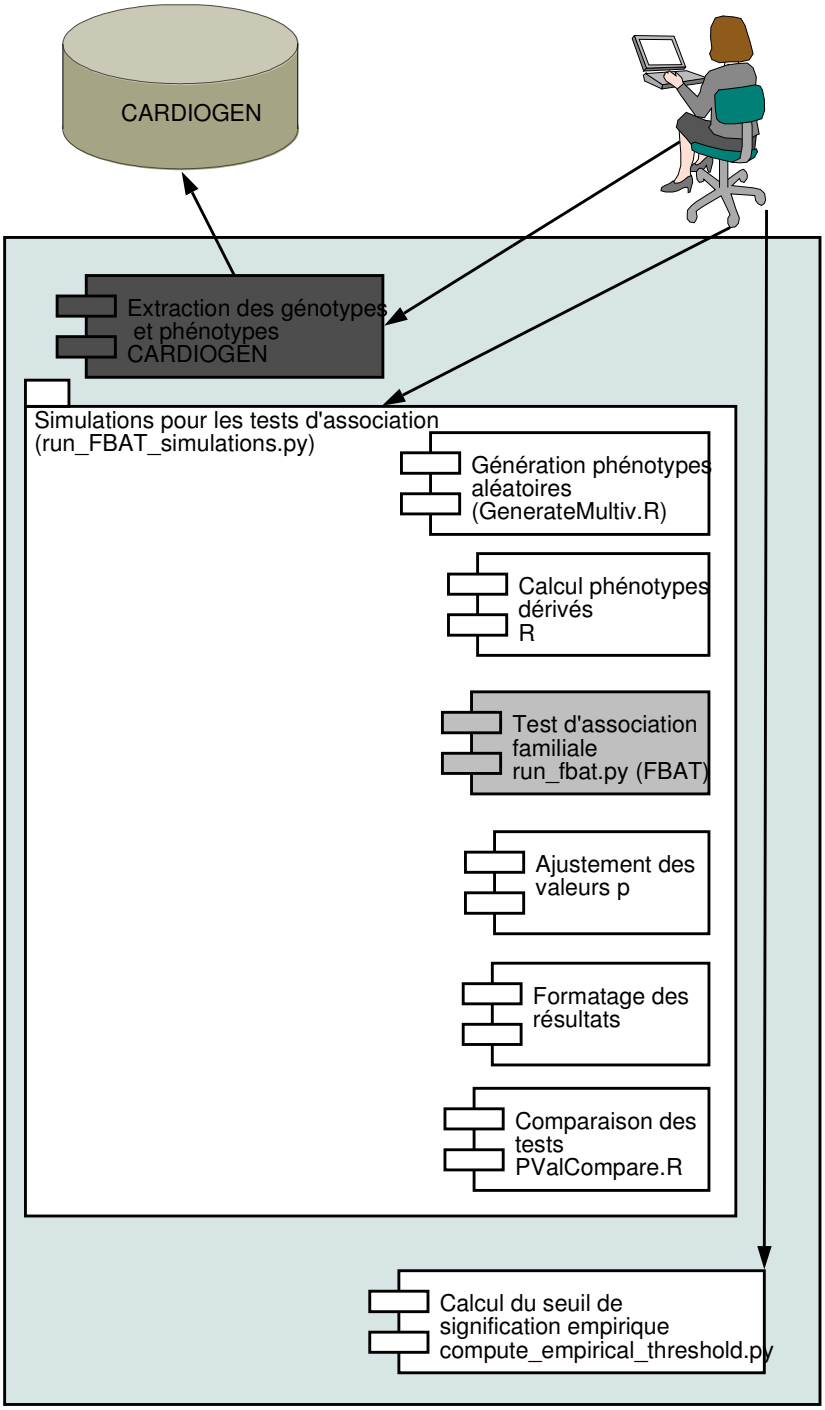


Figure 7.7 Vue de décomposition pour les simulations pour la comparaison des méthodes de réduction de dimensionnalité et le calcul du seuil de signification empirique. La couleur la plus foncée indique que le module existe. La couleur plus claire indique qu'il s'agit d'un nouveau module, la couleur intermédiaire indique une modification de

fonctionnalité existante.

7.4.1 Aspects de l'implantation

Ce mémoire présente deux approches différentes d'implantation qui ont évolué par rapport à l'expérience acquise dans le centre de Recherche du CHUM. Lors de l'implantation de la méthode de réduction de dimensionnalité sur les phénotypes, le principe était plutôt de s'adapter aux logiciels existants et d'adhérer à la méthodologie courante. Dû à cela, l'implantation de la cette partie est restreinte à une application sur un ensemble de phénotypes et ne peut être répétée que dans les conditions expérimentales et en disposant des librairies et packages implantés au Centre de Recherche du CHUM. Il s'agit plutôt de prouver que les techniques d'analyse choisies permettent de trouver des associations significatives tout en contrôlant le taux d'erreurs de type I attendu.

Les analyses avec réduction de dimensionnalité ont été implantées en R et Python. La seule librairie indépendante que nous avons ajoutée est mvtnorm (Genz, et al., 2007).

7.5 Validation statistique des méthodes

Le logiciel développé pour l'analyse d'association repose en grande partie sur des fonctionnalités existantes et sur des logiciels et des librairies externes. Les logiciels externes n'ont pas été validés en soi, mais nous avons vérifié qu'au moins une validation sur les programmes compilés utilisés a été effectuée et documentée. Au CR-CHUM, Brunelle (Brunelle, 2008) a réalisé des tests sur les logiciels Merlin (utilisé dans les simulations pour l'analyse de liaison dynamique), FBAT pour l'analyse d'association familiale et SIBPAL pour l'analyse de liaison. L'incertitude par rapport aux générateurs aléatoires et d'autres propriétés tels que la transmission mendélienne, les distances génétiques, le patron de génotypes manquantes, l'haplotype des fondateurs, le LD et les fréquences alléliques ont été aussi testés ont été vérifiés pour le logiciel Merlin par (Brunelle, 2008).

Les ajustements des valeurs p par rééchantillonnage sont sujets à des fausses découvertes qui proviennent principalement de la méthode de collection des échantillons qui ne représente pas la population générale, l'erreur due à l'utilisation d'un générateur de nombres pseudo-aléatoires et l'erreur du au fait que l'on ne fait pas assez de simulations. Nous avons donc choisi de

réaliser 10000 permutations plus que 1000 lorsqu'il s'agit de confirmer la signification d'un test. Nous avons confirmé par la mesure de l'erreur standard que le nombre de simulations réalisé est adéquat pour l'ajustement de la signification des tests de liaison.

Le générateur de variables aléatoires suivant une loi gaussienne multidimensionnelle a été testée pour vérifier que la corrélation entre les phénotypes est conservée dans les phénotypes simulés. Nous avons choisi les trois premiers ensembles de phénotypes générés et la corrélation entre les phénotypes générés est présentée dans la Figure 7.8. La corrélation des variables est conservée pendant les simulations et nous pouvons faire confiance au générateur aléatoire utilisé. Le test de permutation pour la liaison dynamique utilise aussi un générateur aléatoire, celui de a librairie *random*. Toutes les fonctions de la libraire dépendent de la fonction *random*, qui génère un nombre réel au hasard selon une distribution uniforme dans l'intervalle [0.0, 1.0). Python utilise le générateur de nombres pseudo-aléatoires Mersenne Twister. Il produit des numéros réels distribués uniformément dans 623 dimensions à une précision de 32-bits et a une période de 2^{19937}-1. Le Mersenne Twister est aussi utilisé par les logiciels *Merlin* et *Hapsim*, testés par (Brunelle, 2008). Nous avons testé le générateur indirectement: La liste de corrélations maximales produites pour chaque permutation est enregistrée dans un arbre AVL dont la clé unique est la magnitude de la corrélation. Si le générateur de nombres pseudo-aléatoires ne fonctionnait pas comme attendu, et que les permutations des temps de mesure se répétaient périodiquement dans les simulations, le maximum de la corrélation produite par itération serait identique et la liste de corrélations maximales n'aurait pas après près la même longueur que le nombre de simulations effectuées. Nous contrôlons dans le programme la taille de la distribution nulle et combien de fois une paire phénotype-SNP est associé dans la liste de corrélations maximales.

Des tests unitaires ont aussi été réalisés sur:

- les modifications introduites aux simulations pour la signification des tests de liaison et le test de liaison dynamique par permutation des génotypes (le calcul de la différence des moyennes et la différence des minimums),
- les tests d'association réalisés tant pour la détection de signaux d'association comme pour l'analyse des erreurs de type I et le calcul du seuil de signification par permutations,

- le test de permutation pour la liaison dynamique. Un ensemble de tests a été établi dans la spécification d'exigences du logiciel par rapport aux paramètres, les fichiers d'entrée et le mode d'opération lors d'une défaillance.

Par rapport aux choix méthodologiques, les analyses ont été soumises à consensus dans plusieurs réunions de l'équipe de recherche et, basés sur l'expérience des chercheurs, les méthodes choisies sont appropriées pour tester les hypothèses en question.

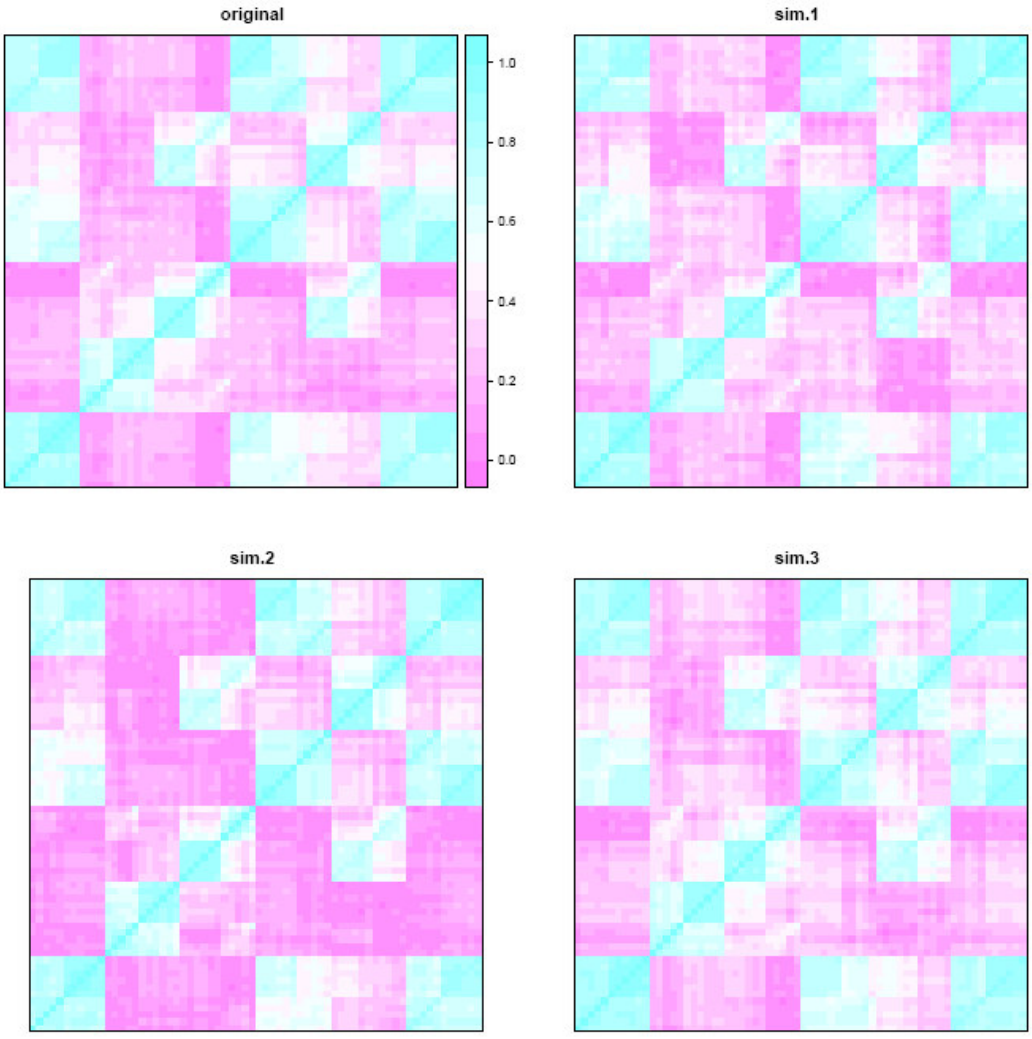


Figure 7.8 Magnitude de la corrélation entre les variables originales et un échantillon des variables simulés par le script GenerateMultiv.R qui utilise la fonction rmvnorm de R. La corrélation entre les phénotypes est conservée pendant les simulations.

CHAPITRE 8. EXPÉRIENCES ET RÉSULTATS

Nous avons disposé de six serveurs de calcul, composés chacun de deux processeurs Intel(R) Xeon à une cadence de 2.40GHz. Les outils ont été testés sur Linux Fedora Core 6 version 2.6.19 noyau 2.6.19-1.2895.fc6 et Python 2.4.4. La version installée de R est la version 2.4. Les résultats du présent chapitre reposent sur les expériences suivantes:

- Permutation pour l'analyse de liaison dynamique entre les valeurs t des tests de liaison (SIBPAL) sur 1000 SNPs et 6 phénotypes mesurés continuellement pendant des tests physiologiques. 10000 permutations ont été effectuées.

- Simulation pour la vérification de la signification des tests de liaison effectuée sur 100 expériences qui ont été choisies selon le niveau de signification observée. Soit qu'ils sont significatifs dans une position et non dans un autre ou simplement significatifs dans les deux positions. La simulation sert à établir la véracité de la liaison observée et à confirmer les résultats des tests de permutation pour la liaison dynamique pour les SNPs qui ont une liaison significative au moins dans une des deux périodes. 10000 permutations ont été effectuées.

- Tests d'association entre 182 individus après élimination des individus avec des phénotypes manquants (sans imputation) et 392 SNPs sélectionnés à partir des tests de liaison. Réduction de dimensionnalité des phénotypes et comparaison des résultats des tests par des tests des différences entre les valeurs p, par des simulations et par une analyse de la distribution des valeurs p par rapport à la distribution attendue selon l'hypothèse nulle. 5000 simulations ont été effectuées.

8.1 Test de permutation pour l'analyse de liaison dynamique

La plupart des marqueurs montrent une variation significative des statistiques de liaison durant les tests physiologiques selon les tests de corrélation, le test indépendant des différences entre les moyennes et le test des différences appariées. Sur un ensemble d'environ 6000 tests, correspondant aux séries de valeurs t des tests de liaison entre 1000 marqueurs et 6

phénotypes, 35% (plus de 2000 couples phénotype-marqueur) sont significativement affectés par les tests physiologiques selon le test de corrélation de Pearson, 29% selon le test de différences des moyennes apparié et 26% selon le test des différences des moyennes sur des variables indépendantes. Après 10000 permutations la proportion de tests significatifs diminue à 26% pour le test de corrélation de Pearson, 23% pour le test apparié et 27% pour le test indépendant. Tel que montré dans la Figure 8.1, selon la distribution nulle empirique, le test déclare comme significatives toutes les corrélations observées plus grandes que 0.95 et dans l'intervalle [0.631, 0.836], à un seuil de signification de 0.05. Pour l'approche des différences des moyennes indépendantes, le test déclare comme significatives toutes les séries dont la statistique t de la différence des moyennes observées est dans l'intervalle [2.306, 4.614]. Finalement, pour l'approche des différences des moyennes appariées, le test déclare comme significatives toutes les séries dont la statistique t de la différence des moyennes observées est dans l'intervalle [2.77, 8.059].

Dans tous les tests un SNP montre une variation significative selon les tests physiologiques pour au moins un phénotype. L'effet des tests physiologiques est donc présent et se manifeste dans tous les marqueurs (997 des mille testés) au moins pour un phénotype. Le calcul des valeurs p empiriques pour les SNPs significatifs dans une position doit nous permettre de confirmer l'aspect dynamique de la liaison ainsi que de confirmer la signification des valeurs p observée pour chaque position. D'ailleurs, il s'avère indispensable de réduire le fardeau des tests multiples pour les tests d'association par position, qui donnent en général des niveaux de signification plus bas que les tests de liaison. La réduction de dimensionnalité des phénotypes s'impose comme alternative pour obtenir des niveaux de signification acceptables séparément pour chaque période.

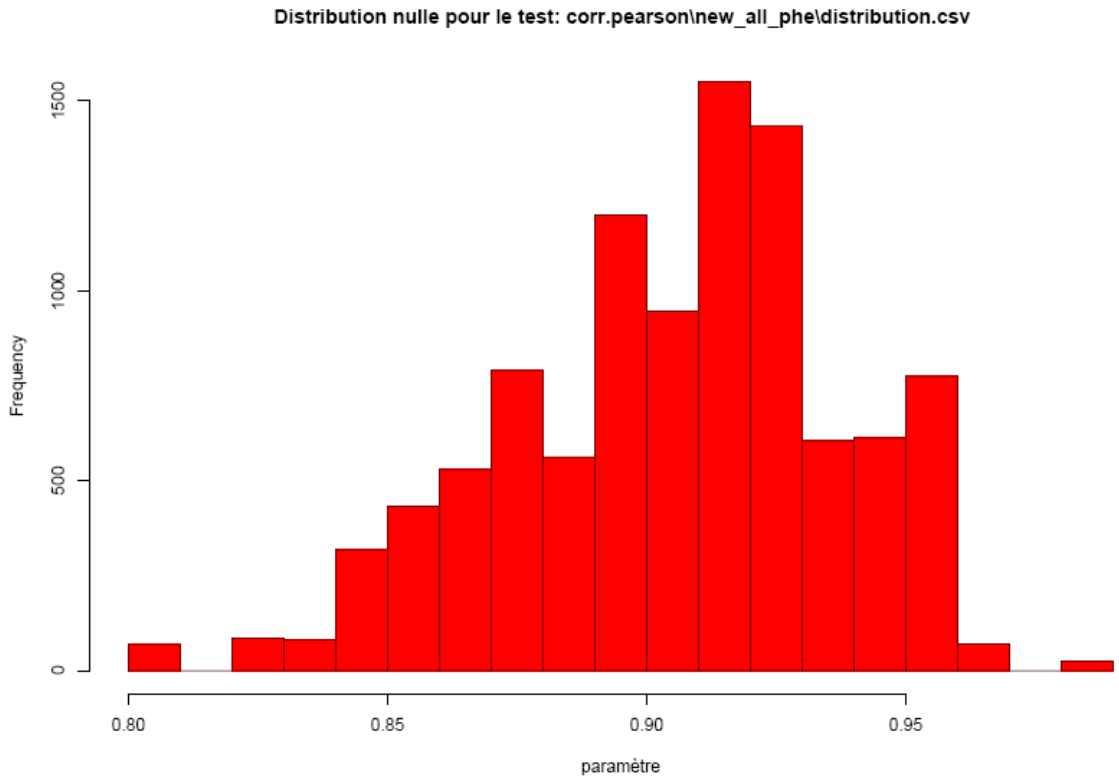


Figure 8.1 Distribution nulle après 10000 permutations pour le test de corrélation de Pearson sur l'ensemble d'environ 6000 expériences.

8.2 Simulations pour la signification de la liaison et de la liaison dynamique

Lors de l'interprétation de résultats des analyses sur les familles dans une population, il est extrêmement utile de savoir combien de fois un résultat similaire pourrait se produire par hasard. Nous calculons combien de fois la valeur t (la statistique du test) minimale est supérieure à la statistique observée pour chaque position. La fréquence à laquelle cet événement se produit devient la nouvelle valeur p du test de liaison par période. Pour le test de liaison dynamique nous calculons combien de fois une statistique (la différence des minimums des deux périodes et la valeur t du test de différences des moyennes entre les deux périodes) dans les permutations est inférieure à la valeur observée de la même statistique. Le nombre de simulations à effectuer doit être généralement lié au seuil de signification choisi pour l'expérience. Plus la valeur de N est grande, plus les valeurs critiques seront calculées sur des

estimations précises. Pour obtenir des estimations plus stables nous avons décidé de fixer N à 10000. Sur ces 10000 nouveaux ensembles statistiques du test de liaison, deux tests ont été retenus: un test sur les différences des minimums des deux périodes et un test de différences de moyennes par position. 100 combinaisons phénotype-génotype ont été testées par ce procédé et les résultats de ces tests son rapportés dans le Tableau A.1 (annexe 1).

Pour toutes les expériences testées, la liaison est significative après 10000 permutations pour au moins une des périodes. Quelques marqueurs dans des régions chromosomiques ayant été rapportées dans la littérature ont été confirmés par cette méthode. Particulièrement nous attirons l'attention sur les SNPs dans les chromosomes 1 et 17 qui sont en proximité avec des marqueurs signalés dans des études des phénotypes liés à la pression sanguine sur l'étude Framingham, une des plus grandes initiatives existantes pour dévoiler les facteurs de risque des maladies cardiovasculaires. Le grand nombre de marqueurs significatifs nous ouvre aussi une voie intéressante sur des régions qui restent jusqu'à maintenant inconnues.

Le Tableau A.1 montre quelques régions intéressantes soit parce qu'elles sont significativement en liaison génétique ou parce que le gène présent dans la région est significativement associé à une des phénotypes choisis dans l'étude et que la fonction du gène a été établie et répertoriée dans l'outil NCBI maps.

La plupart de marqueurs qui sont liés dans une période, le sont dans la période couché. Il semblerait que lorsque la variabilité environnementale est grande, les tests deviennent significatifs. Lorsque les tests sont significatifs dans les deux positions, c'est encore dans la position couché que les paires SNP phénotype obtiennent des valeurs p plus petites, impliquant que la relation entre les phénotypes et les génotypes semble être plus "extrême" dans la position couché, ou plus vrai que ce qui aurait pu l'être juste par hasard. Nous pouvons lancer l'hypothèse que dans la position couchée l'influence de l'environnement est moindre que dans la position debout (les individus sont plus influencés par des stimuli extérieurs), où les variations environnementales prennent l'avance sur les variations génétiques.

Les résultats des deux tests retenus pour l'analyse de liaison dynamique ne sont pas moins intéressants: Lorsque les marqueurs sont liés dans la position couché, le test de liaison dynamique réussi à 85% à détecter une différence significative. Par contre, lorsque les tests sont significatifs dans la position débout le test de liaison dynamique ne détecte que la moitié des fois la différence. En analysant à profondeur, 4 fois sur 8 les tests ont une valeur p de l'ordre

de 1x10-E02, très près du seuil de signification de 5x10-E02. Les quatre autres cas restent incertains. C'est intéressant aussi de souligner que lorsque les SNPs sont liés dans les deux positions, le test de liaison dynamique ne détecte pas de différence le 84% des fois. Le restant 16% corresponde à des tests de liaison dont les valeurs p diffèrent de 1/100. Le deuxième test de liaison dynamique vérifie combien de fois la valeur t du test de différence des moyennes entre les valeurs t des deux périodes est plus grande en magnitude dans les simulations que dans les données observés. Nous avons pu observer que la différence des moyennes est beaucoup plus variable que la différence des minimums. Cela nous permet de conclure que n'importe quelle statistique moins stricte que le minimum pour calculer la signification de la liaison par période serait plus sensible aux variations des simulations point par point.

10000 simulations ont été réalisées pour chaque paire SNP-génotype. Nous avons calculé l'erreur standard de la valeur p ajustée étant donnée le nombre de permutations. Si l'intervalle de confiance a 95% après 10000 permutations comprend le seuil α=0.05, les simulations devraient se poursuivre. Nous avons calculé les intervalles de confiance pour les tests de liaison dynamique et pour le test de signification. 3 sur 100 tests dans la position couché devraient se poursuivre selon cette estimation. Nous avons décidé de ne plus faire des simulations car pour ces SNPs la liaison est aussi significative dans la position debout.

Tableau 8.1 SNPs en liaison dynamique localisés sur des gènes et sa fonction associée tirée de NCBI maps.

SNP	***Phénotype***	***Chr.***	***Gène***	***Fonction***
rs10489184	dia	1	F5 coagulation factor V	Coagulation du sang
rs3753396	sys	1	CFH complement factor H	Fonction rénale et oculaire, associé à la susceptibilité è l'infarctus du myocarde
rs395404	sv	2	SLC8A1 solute carrier family 8 (sodium/calcium exchanger),	Régulation de la force de la contraction du cœur.
rs10513688, rs10513689	dia, map,	3	SLC2A2 solute carrier family 2	Diabètes mellitus, non dépendent à l'insuline- les problèmes cardiovasculaires et l'hypertension sont des complications associées au diabète
rs7661217	tpr	4	CORIN corin serine peptidase	La protéine encodée active une hormone cardiaque qui régule le volume du sang et la pression sanguine

SNP	*Phénotype*	*Chr.*	*Gène*	*Fonction*
rs10519945 rs10519959 rs3846329 rs4835136	map, sys sys, dia dia sys, dia	4	NR3C2 nuclear receptor subfamily 3, group C, member 2	Associé au risque de début précoce de l'Hypertension, avec aggravation dans la grossesse
rs7759088	map, sys	6	PLN phospholamban	Circulation du sang, développement du muscle cardiaque, régulation de la contraction du cœur et de la force de la contraction du cœur.
rs6961069	tpr	7	CD36 molecule (thrombospondin receptor)	Associé à la susceptibilité de la maladie coronarienne du cœur.
rs10514919	dia, map	17	ITGB3 integrin, beta 3	Associé à la coagulation du sang et proche à la région rapporté par Levy sur le Framingham heart study
rs1924921	tpr	13	EDNRB endothelin receptor type B	Régulation de la pression sanguine
rs10512510	sv	17	PRKCA protein kinase C, alpha s	Régulation de la force de la contraction du cœur. Études sur des souris suggèrent que le produit du gène est un régulateur fondamental de la contractilité cardiaque et du contrôle de Ca(2+) dans les myocytes (cellules musculaires).
rs1799898	tpr	19	LDLR low density lipoprotein receptor	Des mutations dans ce gène causent l'hypercholestérolémie familiale (maladie qui peut engendrer des complications cardiovasculaires).

Sur les 100 simulations réalisées, 72% ont été bien classifiées selon le Test de Permutation sur la corrélation, 60% sur les différences indépendantes et 55% sur le test des différences appariés. Le test sur les différences des moyennes appariées est très conservateur par rapport aux deux autres et il semble ne pas être nécessaire de sacrifier la signification en faveur de la détection des différences lorsque le signal de liaison contient des pics de signification. La valeur p du test de différences entre les minimums des valeurs t obtenues en simulant les génotypes est plus relevant du point de vue génétique parce que le résultat est conditionnel aux phénotypes et aux génotypes observés.

8.3 Analyse d'association familiale avec réduction de dimensionnalité des phénotypes

En total nous comptons avec 55 variables reliés à la pression sanguine. Ces variables correspondent à 5 phénotypes mesurés de façon répétée à chaque 5 minutes pendant la position couchée et à chaque 2 minutes pendant la position debout. Nous avons réalisé une analyse de composantes principales et, selon la règle de "valeur propre plus grande que 1", nous devrions prendre au moins les 8 premières composantes. Cela confirme que la variabilité est différente dans les deux périodes, car en séparant les variables par période, l'analyse de composantes principales et la même règle de sélection nous fournissent 4 composantes principales par période. Nous en parlerons plus précisément dans la section de l'ACP.

8.3.1 Réduction par la méthode des moyennes

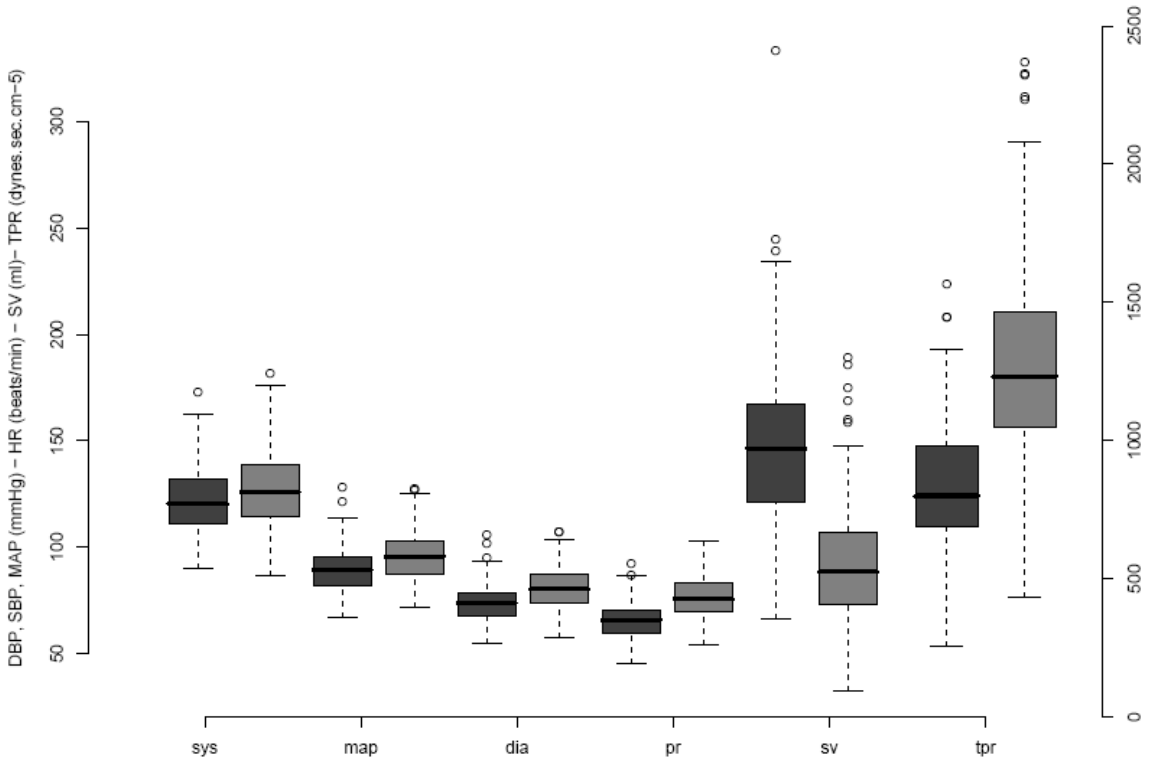
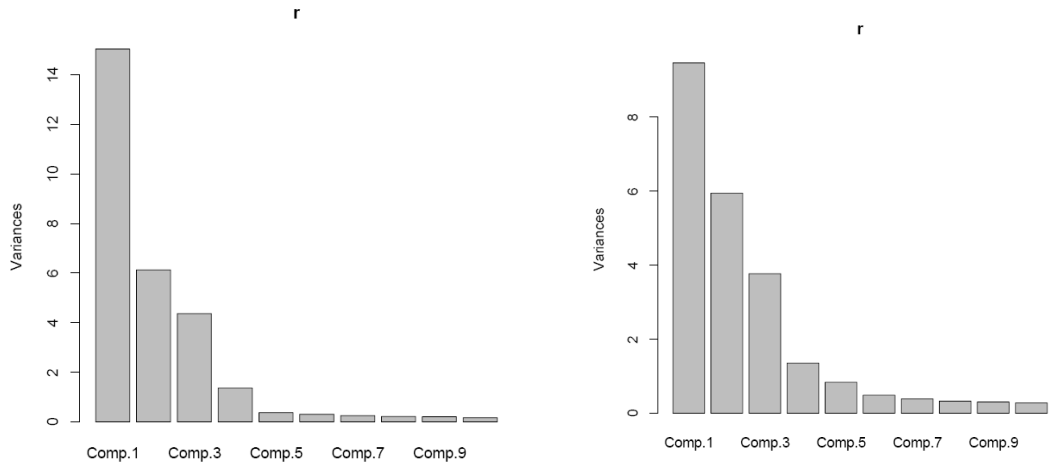


Figure 8.2 Comparaison entre les phénotypes créés à partir de la méthode de la moyenne. Les barres foncées équivalent à la position couchée et les barres claires à la position debout. L'unité de mesure sur le côté droit du graphique s'applique au phénotype TPR.

Lorsque les mesures répétées de chaque phénotype sont réduites à la moyenne par chaque période, on constate que tous les phénotypes montrent des changements dramatiques d'une position à l'autre (les valeurs p des tests différences appariées pour tous les phénotypes étant

plus petites ou égales que 1*10-13) ainsi qu'une augmentation de la variance dans les phénotypes SBP, SV et TPR. Les phénotypes "moyenne" reflètent bien la variation par rapport aux tests physiologiques, et ce fait peut être constaté dans la Figure 8.2.

8.3.2 Réduction par l'ACP



(a) (b)

Figure 8.3 Valeurs propres des composantes obtenues à partir de l'ACP dans les positions couché (a) et debout (b). On retient les 4 premières composantes

Les phénotypes intermédiaires de la pression sanguine sont exprimés en différentes unités. Par le fait même, ils ont été standardisés avant d'être ajoutés à l'ACP, cette standardisation consistant à centrer chaque mesure par rapport à sa moyenne et diviser pour l'écart type de tous les individus (séparément pour chaque mesure dans le temps). Pour les positions debout et couchée, les 4 premières composantes expliquant 90% et 83% de la variance phénotypique ont été retenues (voir la Figure 8.3). La différence des moyennes d'une période à l'autre disparaît par le fait que l'on centre et réduit les phénotypes originaux et ce fait se constate dans la Figure 8.4.

Chaque composante encapsule la variabilité d'un ou plusieurs phénotypes et sa contribution sera plutôt reliée à la diminution du nombre de tests effectués en faisant que l'ajustement par tests multiples soit moins rigoureux. La Figure 8.5 montre les facteurs de pondération (corrélation entre les variables originales et les nouvelles composantes) pour l'analyse des composantes principales dans la position debout. Les variables sont projetées sur le plan des 4

composantes choisies. Clairement on voit un effet de regroupement des variables correspondant à des mesures répétées. Dans la position couchée, la magnitude des facteurs de pondération des phénotypes TPR et SV est plus grande sur la composante 1. Celle de la pression systolique est plus grande sur la composante 2, le pouls à sa fois sera apparié avec la composante 3 et la pression diastolique avec la composante 4. Dans la position debout, la magnitude des facteurs de pondération pour les phénotypes TPR et SV est plus grande sur la composante 1. Celle des facteurs de pondération pour la pression systolique (SYS) sur la composante 2, le pouls (HR) sur la composante 3 et la pression diastolique (DIA) sur la composante 4. Sur tous les plans et dans les deux positions les variables HR et SV sont orthogonales par rapport à SYS et DIA, ainsi que TPR par rapport à SYS et DIA. Ce fait nous permet de dire que dans le contexte de l'analyse les variables HR, SV et TPR sont indépendants des pressions artérielles SYS et DIA.

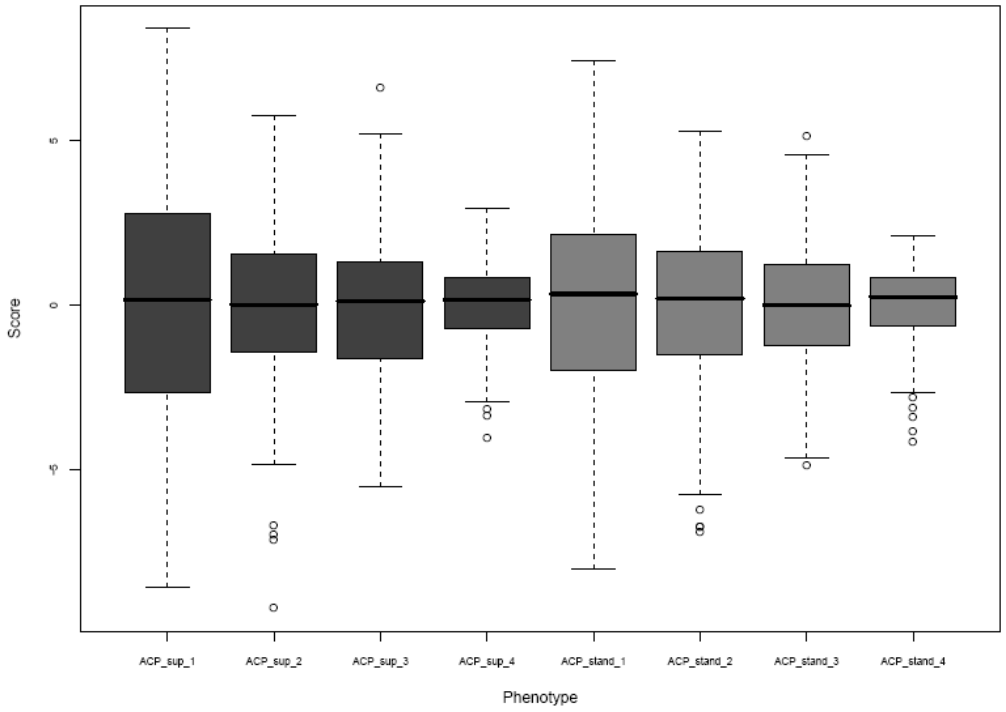
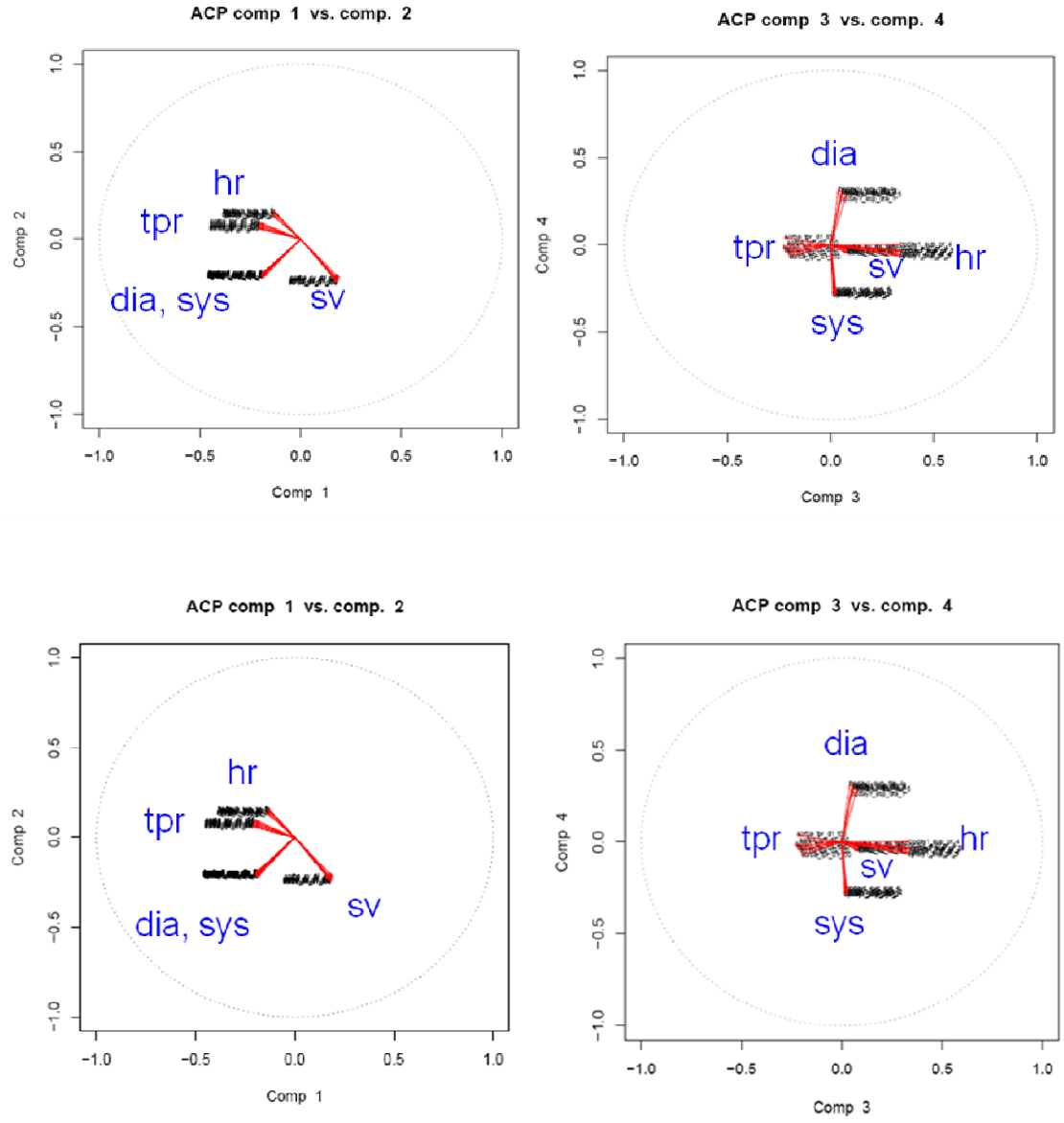


Figure 8.4 Comparaison entre les distributions des phénotypes créés à partir de l'ACP. Les barres foncées équivalent à la position couchée et les barres plus claires à la position debout.



(b)

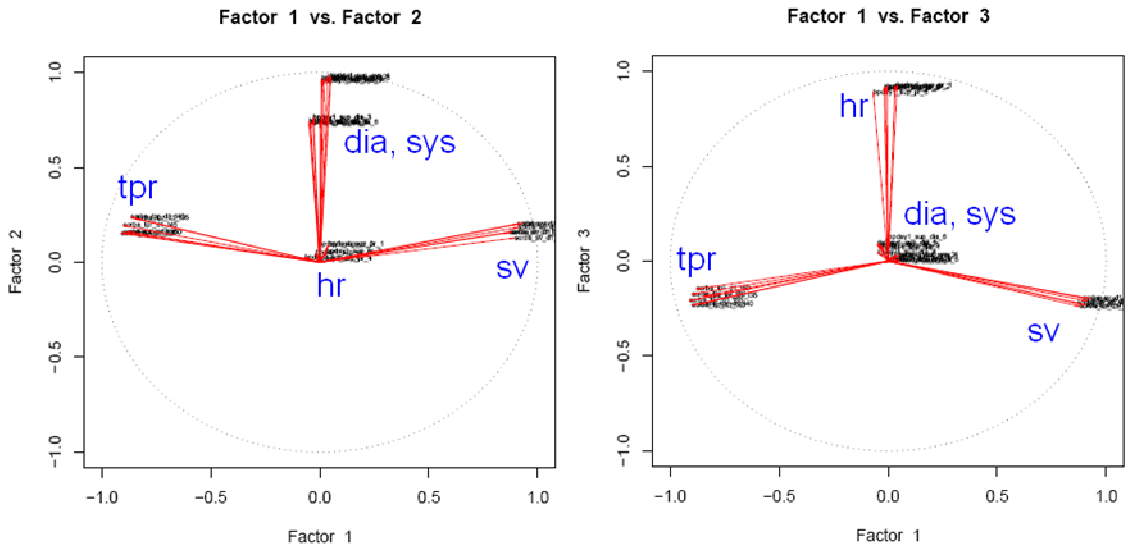
Figure 8.5 Variables originales projetés dans l'espace des quatre premières composantes de l'ACP pour les positions couché (a) et debout (b)

8.3.3 Réduction par l'AF

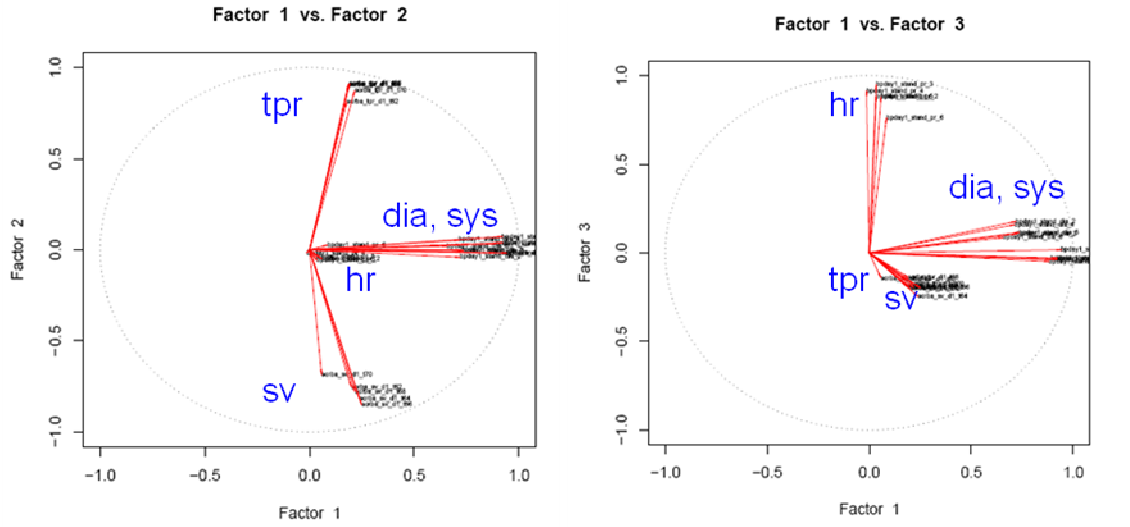
À l'égal que dans la réduction par l'ACP, les phénotypes ont été standardisés avant d'être ajoutés au modèle. Pour les positions debout et couchée, les 3 premiers facteurs expliquant

95% et 89% de la variance phénotypique ont été retenus par position. Similairement, la différence des moyennes dans ces nouveaux phénotypes d'une période à l'autre n'est pas significative. Chaque facteur contribue à la diminution du nombre de tests effectués en faisant que l'ajustement par tests multiples soit moins rigoureux.

La Figure 8.6 montre les variables projetées dans l'espace des trois premiers facteurs par position. Pour que chaque une relation univoque existe entre la variable original et un seul facteur nous avons utilisé des rotations orthogonales. La rotation varimax est celle qui regroupe mieux les phénotypes selon ce qui est attendu: la pression systolique et diastolique ensemble et les autres phénotypes indépendants de ces deux. Les nouvelles variables restent indépendantes après l'application de la rotation. L'effet de regroupement des variables est aussi présent et l'appariement « un-phénotype-un-facteur» est encore une fois faisable. Dans la position couchée, les valeurs p reliées aux phénotypes TPR et SV sont appariées avec les résultats du facteur 1. Celles des pressions systolique et diastolique à ceux du facteur 2, le pouls étant apparié avec le facteur 3. Dans la position debout, le facteur 1 est apparié avec les pressions systolique et diastolique, le facteur 2 avec le TPR et le SV et le facteur 3 avec le pouls.



(a)



(b)

Figure 8.6 Les facteurs de pondération de l'AF pour les positions couché (a) et debout (b) (variables projetées dans l'espace des trois facteurs sélectionnés).

Sur tous les plans et dans les deux positions les variables SV et TPR sont approximativement orthogonales par rapport à SYS et DIA, ainsi que HR par rapport à SYS et DIA. Tel que nous l'avons notée dans l'analyse de composantes principales les variables HR, SV et TPR sont indépendants des pressions artérielles SYS et DIA.

8.3.4 Ajustement par la méthode de Bonferroni

Tel que montré dans la Figure 8.7, l'ajustement par la méthode de Bonferroni consiste à ajuster les valeurs p de tous les points d'un essai particulier. Le minimum de chaque période est multiplié par le nombre d'essais de chaque période. Dans le cas particulier de la figure ci-dessus, le test reste significatif pour la période couchée et non significative pour la période debout. Ce faisant, nous n'assurons pas un contrôle fort du taux de faux positifs. Un contrôle rigoureux tiendrait compte du fait que l'on teste à la fois plusieurs SNPs.

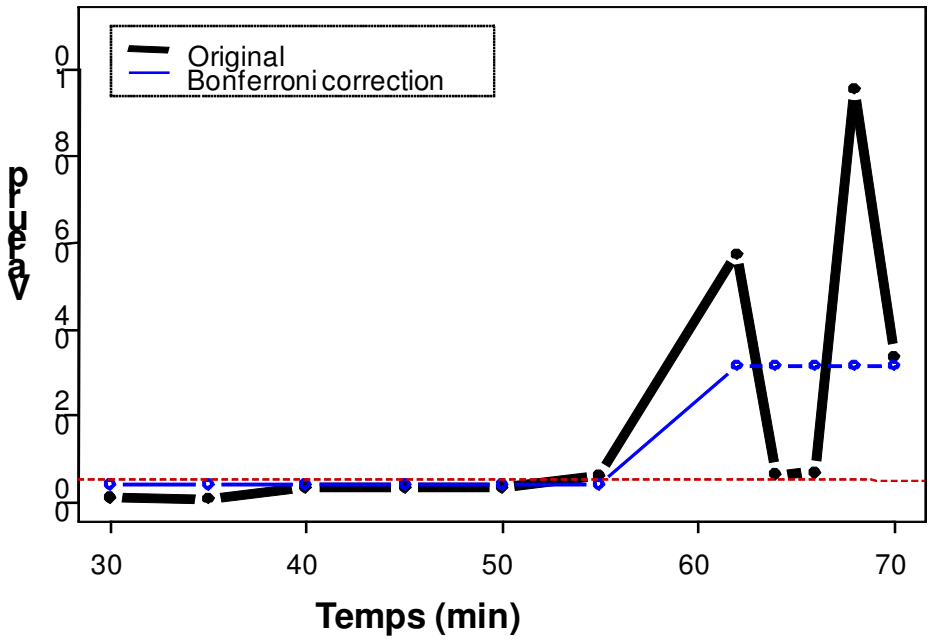


Figure 8.7 Valeurs p originales et ajustées pour le phénotype HR, SNP rs1855615. La ligne pointillée défini le seuil de signification (0.05)

Tel qu'attendu, la méthode de Bonferroni est la correction la plus rigoureuse des valeurs p pour l'ensemble des tests d'association. Dans les sections suivantes nous analysons les différences dues à l'application des méthodes de réduction et nous comparons les valeurs p obtenues avec celles corrigées par la méthode de Bonferroni.

8.3.5 Comparaison des méthodes

Le Tableau 8.2 montre les résultats des tests sur les différences des valeurs p obtenues par les quatre méthodes. On peut constater que n'importe quelle méthode est meilleure que la correction par la méthode de Bonferroni et que les valeurs p ne sont pas significativement différentes par les autres trois méthodes, tous les phénotypes confondus. D'autres méthodes s'imposent pour faire ressortir la méthode la plus efficace.

On compare la quantité d'essais significatifs par période par méthode et on obtient les résultats de la Figure 8.8. On peut bien dire que l'ACP et l'AF ont le plus grand nombre d'essais significatifs et que dans l'ensemble le taux d'erreurs de type I se conserve bien (même si le risque existe que tous les résultats rapportés soient des faux positifs à un seuil $\alpha = 0.05$).

Tableau 8.2 Résultats des tests des différences appariés pour la comparaison des valeurs p des tests d'association (*** = p < 1 e-06)

Hypothèse alternative: les valeurs p du modèle à gauche sont plus petites que ceux du modèle à droite		**période**	**significative**
p.FA	p.ACP	couché	Non
p.FA	p.aver		Non
p.ACP	p.aver		Non
p.aver	p.Bonf		***
p.ACP	p.Bonf		***
p.FA	p.Bonf		***
p.ACP	p.aver	debout	Non
p.FA	p.ACP		Non
p.FA	p.aver		Non
p.aver	p.Bonf		***
p.FA	p.Bonf		***
p.ACP	p.Bonf		***

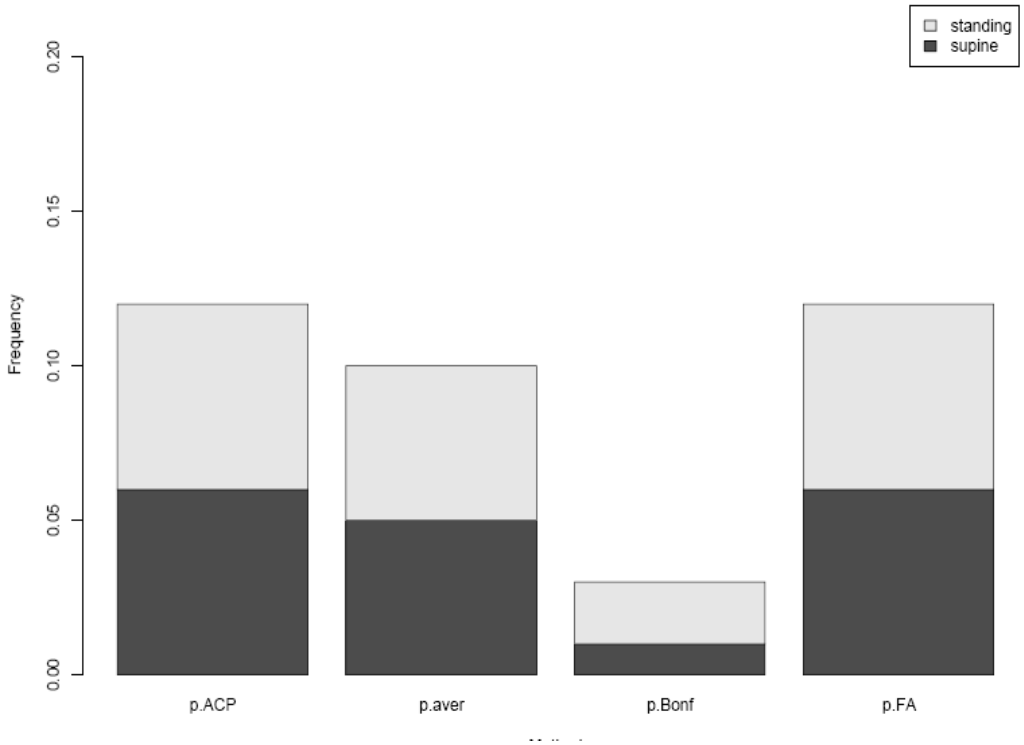


Figure 8.8 Proportion d'essais significatifs à un seuil α=0.05 pour l'analyse d'association familiale avec réduction de dimensionnalité des phénotypes

Le graphique log quantile-quantile des valeurs p est très utile pour interpréter les résultats des tests où plusieurs SNPs sont compromis. Une analyse de la distribution des valeurs p selon la méthode utilisée a été réalisée. Il s'agit donc de comparer les valeurs p obtenues par chaque méthode par rapport à ce qu'on s'attendait selon l'hypothèse nulle. Sous l'hypothèse nulle de non-association, on espère que chaque essai ait la même probabilité d'être significatif et que les valeurs p soient en conséquence distribuées uniformément dans l'intervalle [0,1]. Nous avons dessiné la distribution des valeurs p observées par chaque modèle par rapport à la distribution attendue et la Figure 8.9 a été obtenue. L'adhérence des valeurs p à la plus part de la distribution théorique montre qu'il n'y a pas trop de sources d'associations aberrantes. Bonferroni est très conservateur, les autres méthodes se ressemblent sauf pour la plus grande tendance de l'ACP vers des valeurs p plus grandes. L'ACP est donc choisi comme la méthode qui donne les plus grands niveaux de signification sans pour tant s'éloigner de la distribution attendue selon l'hypothèse nulle.

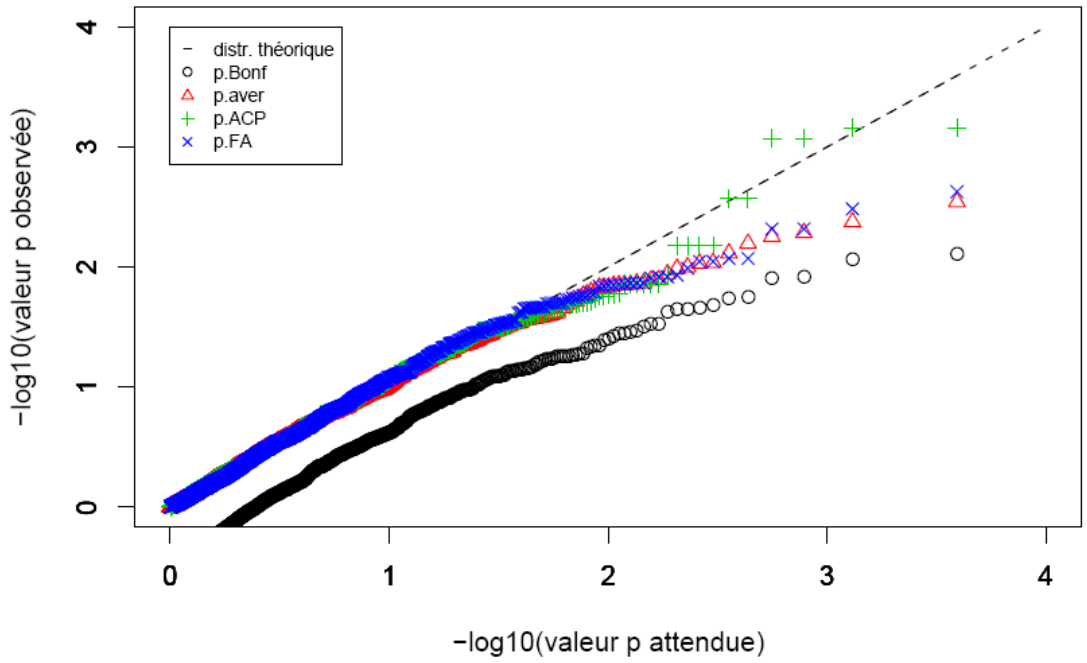


Figure 8.9 Diagramme log quantile-quantile pour les valeurs p des tests d'association de 392 SNPs et 5 phénotypes avec les différentes méthodes. L'adhérence des valeurs p à la plus part de la distribution théorique montre qu'il n'y a pas trop de sources d'associations aberrantes.

5000 itérations des simulations expliquées dans la section 6.5.1 ont permis de confirmer que le niveau de signification se préserve bien dans les trois méthodes, la méthode de composantes principales est celle qui a le taux le plus haut d'erreurs de type I pendant les 5000 itérations (Voir la Figure 8.10).

Nous nous sommes servis des permutations pour estimer l'erreur de type I pour notre ensemble de données particulier. Après 5000 permutations pour la méthode de composantes principales on pourrait rejeter l'hypothèse nulle non-association en présence de liaison pour des valeurs p plus petites que 0.00134 dans la position couchée et 0.00136 dans la position debout. Les seuils de signification ajustés pour les méthodes de la moyenne et l'analyse factorielle sont très semblables (de l'ordre de 10^{E}-03). Nous présenterons maintenant l'ensemble de marqueurs dont la valeur p nous permet de rejeter l'hypothèse nulle aux seuils de signification obtenus par simulation.

Tableau 8.3 Seuil de signification calculé à partir de permutations du phénotype. Les seuils ajustés sont très semblables pour les trois méthodes de réduction.

Méthode	*Période*	*Moyenne du maximum de la statistique obtenue par simulation*	*Quantile 95% sur le maximum de la statistique obtenue par simulation*	*Seuil ajusté*
MOY	couché	2,810	3,264	1,10E-03
MOY	debout	2,815	3,243	1,18E-03
AF	couché	2,709	3,168	1,53E-03
AF	debout	2,707	3,152	1,62E-03
ACP	couché	2,774	3,207	1,34E-03
ACP	debout	2,774	3,204	1,36E-03

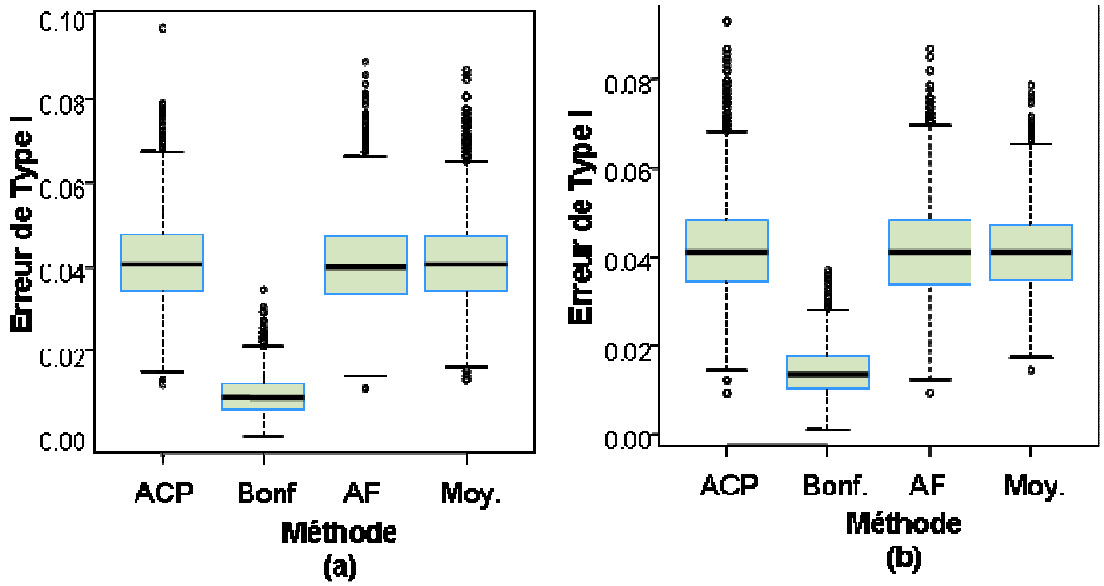


Figure 8.10 Erreur de Type I pour les tests d'association dans les études de simulation pour les positions couchée (a) et debout (b)

Le tableau 8.4 montre les marqueurs ayant été détectés comme associés ou faiblement associés. Trois marqueurs situés sur le gène C10orf59 dans le chromosome 10 sont ressortis avec un niveau de signification de 6.95E-04, 8.56E-04 et 6.6310E-03 dans la position couché. Une association faiblement significative est aussi détectée dans le même marqueur par la

méthode d'analyse factorielle avec une valeur p de 2.35 E-03. Le gène C10orf59 est associé à la Renalase, protéine secrété dans le sang par le rein et reconnue parce qu'elle module la fonction cardiaque et la pression sanguine. Ce résultat confirme l'évidence d'association rapporté dans la population Han en Chine. La découverte de cette association est une voie pour dévoiler les mécanismes de régulation de la pression systolique et la pathogénèse de l'hypertension essentielle. Le SNP rs1855615 dans le chromosome 9 a aussi une association avec les phénotypes HR et TPR dont la valeur p est à peine proche du seuil de signification calculé avec une valeur p de 2.6 E-03 dans la position couché. Cette association est aussi détectée par l'analyse factorielle (2.36E-03). Ce SNP est localisé sur le gène TEK (tyrosine kinase, endothélial) qui code la protéine TEK (*receptor tyrosine kinase)* reconnue par son effet dans la malformation des veines. Nous portons aussi l'attention sur le SNP rs721619 dans le chromosome 8 ayant un signal d'association avec SV et HR dans la position debout de l'ordre de 10E-03 dans tous les méthodes (moyenne, analyse factorielle et Bonferroni) sauf l'ACP. Ce SNP est localisé sur le gène EPHX2 qui a été rapporté dans la littérature comme associé au risque d'incidence de CHD (maladie coronarien) dans des populations caucasiens et indiquent que le gène est un gène de susceptibilité des maladies cardiovasculaires (ainsi relié dans la population chinoise comme gène de risque pour l'accidente cérébrale vasculaire).

Tableau 8.4 SNPs significatifs dans les tests d'association et leur localisation dans les chromosomes.

Méthode	*Valeur p*	*Rs ID*	*Chr*	*Gène*	*Link*	*Position*
ACP	6.60E-03	rs10509549	10	C10orf59	http://www.ncbi.nlm.nih.gov/projects/mapview/maps.cgi?taxid=9606&chr=10&MAPS=ugHs,genes,snp-r&cmd=focus&fill=40&query=uid(119952682)&QSTR=rs10509549	90204973
ACP	6.95E-04	rs10509547	10	C10orf59	http://www.ncbi.nlm.nih.gov/projects/mapview/maps.cgi?taxid=9606&chr=10&MAPS=ugHs,genes,snp-r&cmd=focus&fill=40&query=uid(119952680)&QSTR=rs10509547	90202230
ACP	8.56E-04	rs10509548	10	C10orf59	http://www.ncbi.nlm.nih.gov/projects/mapview/maps.cgi?taxid=9606&chr=10&MAPS=ugHs,genes,snp-r&cmd=focus&fill=40&query=uid(119952681)&QSTR=rs10509548	90203442
ACP	2.68E-03	rs1855615	9	TEK tyrosine kinase, endothelial	http://www.ncbi.nlm.nih.gov/projects/mapview/maps.cgi?TAXID=9606&CHR=9&MAPS=ideogr%2Ccntg-r%2CugHs%2Cgenes&QUERY=27211874&BEG=27211870&END=27211879&thmb=on	27211874
AF, Bonferroni, moyenne	3.28E-03, 8.47E-03, 5.17E-03	rs721619	8	EPHX2	http://www.ncbi.nlm.nih.gov/projects/mapview/maps.cgi?taxid=9606&chr=8&MAPS=ugHs,genes,snp-r&cmd=focus&fill=80&query=uid(132281057)&QSTR=rs721619	27437913

CHAPITRE 9. CONCLUSION ET DISCUSSION

9.1 Rappel des expériences réalisées et résultats significatifs

Nous avons conçu, implanté et réalisé une analyse de liaison dynamique sur un ensemble marqueurs génétiques et des mesures répétées de phénotypes associées à la pression artérielle. Nous nous sommes d'abord engagées dans une analyse de la variation des signaux de liaison génétique pendant le temps pour confirmer que les variations répondent aux tests physiologiques. Ensuite nous avons réalisé des simulations pour confirmer la signification des tests dans les périodes définis par les tests physiologiques et pour mesurer la signification de la liaison dynamique en tenant compte des phénotypes, des génotypes, de la distribution des donnés manquantes et de la généalogie. Quelques marqueurs dans des régions chromosomiques ayant été déjà rapportés dans la littérature ont été confirmés par cette méthode. D'autres marqueurs dans des gènes dont la fonction n'a pas été établie sont aussi significatifs après 10000 permutations.

Sur un ensemble de SNPs sélectionnés à partir des tests de liaison nous avons implanté et réalisé des tests d'association familiale. Nous avons utilisé trois techniques différentes de réduction de dimensionnalité des phénotypes par chaque période et nous avons comparé l'effet de chaque méthode en confrontant directement les valeurs p des tests d'association selon chaque technique. Nous avons par la suite implanté des simulations pour confirmer que ces techniques contrôlent le taux d'erreurs de type I et pour établir des seuils de signification pour chaque méthode. La similarité générale entre les méthodes indique qu'il n'y a pas une perte significative d'information en réduisant les multiples mesures des différents phénotypes. On peut se fier aux résultats fournis pour l'ACP, en se basant sur les seuils de signification obtenus par simulation. Les réductions de dimensionnalité fournies par l'ACP offrent une bonne représentation de la variance phénotypique et de la corrélation entre les variables, en préservant le taux d'erreur de type I attendu. L'ACP a fourni quelques associations significatives qui n'auraient pas été découvertes par des méthodes d'ajustements traditionnels tels que la méthode de Bonferroni. L'ACP est aussi une méthode de choix si l'on ne peut pas assurer que les données proviennent d'une distribution gaussienne multidimensionnelle.

L'expérience des simulations pour la signification des tests de liaison nous a permis d'obtenir un ensemble important de marqueurs génétiques significativement liés aux traits intermédiaires de la pression artérielle et il est maintenant clair que nous aurions pu obtenir plus de marqueurs associés à nos traits corrélés sur des marqueurs en déséquilibre de liaison en utilisant une méthode de rééchantillonnage. (Brunelle, 2008) a implanté une méthode d'ajustement des valeurs p des tests d'association familiale par rééchantillonnage qui tient compte du fait que les phénotypes sont corrélés et du fait que les marqueurs peuvent être en déséquilibre de liaison. Cette méthode serait plus puissante pour la détection d'associations que notre méthode de réduction de dimensionnalité combinée avec la relaxation du seuil de signification par simulations.

9.2 Limitations

Nous avons réalisé plusieurs analyses qui utilisent les méthodes de rééchantillonnage. Nous avons utilisé une méthode de permutation pour quantifier la signification de l'effet des tests physiologiques par rapport aux marqueurs disponibles, la généalogie et les phénotypes mesurés. Cette méthode procède en simulant les génotypes qui respectent quelques propriétés des génotypes originales. Mais les simulations procèdent indépendamment par chaque paire SNP-Phénotype. Cela implique que nous ne corrigeons pas par le fait que nous testons sur plusieurs génotypes en LD.

Nous avons aussi utilisé des simulations pour comparer les méthodes de réduction de dimensionnalité des phénotypes pour tester l'association entre les phénotypes intermédiaires de l'hypertension et l'ensemble des marqueurs significativement liés détectés dans la premier partie de l'étude. Nous avons choisi d'échantillonner les phénotypes à partir d'une distribution gaussienne multidimensionnelle pour assurer que les phénotypes simulés reproduisent la distribution des phénotypes originaux. En ce faisant, nous avons détruit la variabilité intrafamiliale et notre simulation s'avère moins exacte par rapport à une méthode comme celle utilisé par PLINK ou les phénotypes sont permutés à l'intérieur des familles. Voici une voie qui s'ouvre pour des travaux postérieurs.

Finalement, notre approche manque aussi de validation par rapport aux logiciels produits. Nous avons réduit la probabilité d'erreurs due à l'intégration de logiciels existantes en vérifiant que des tests unitaires e d'intégration sur les logiciels existantes ont été effectués. Nous avons aussi

validé les caractéristiques des générateurs de nombres pseudo-aléatoires que nous avons utilisés. Nous avons réalisé des tests unitaires sur tous les logiciels produits, mais tous les tests ont été effectués sur un seul jeu de données.

Ensuite, nous voulons principalement rappeler un problème très commun dans l'épidémiologie génétique et apporter quelques suggestions de solution. Les équipes de recherche en général sont multidisciplinaires et sont dirigées principalement par des gens avec peu de formation en génie logiciel. Étant sans doute un aspect important dans les découvertes scientifiques, les étapes de production des logiciels associés à la découverte ne suivent pas nécessairement des normes standard. En général, on accuse le manque de temps pour l'établissement des besoins, le design, les validations et les tests de conformité des logiciels produits. Nous avons essayé avec le test de permutation pour la liaison dynamique d'établir une culture de conscientisation sur les étapes du cycle de vie de notre logiciel en produisant un document de spécification de besoins accordé au standard IEEE. Mais malheureusement nous n'avons pas reçu de rétroaction sur la documentation produite. Nous avons réalisé des tests unitaires sur les logiciels, mais la validation de conformité du produit avec les besoins est faite par rapport à l'a conformité du choix de la méthode par rapport au problème proposé. Et en absence d'un expert dans le domaine biostatistique des erreurs de conception peuvent découler du choix des méthodes ainsi que du manque de formation biologique des informaticiens.

Un pont vital entre le langage du génie logiciel et le langage médical ou biologique est fait par les bioinformaticiens. La bioinformatique étant bi-disciplinaire, la formation du cycle de vie du logiciel n'est jamais suffisante, et dans quelques cas ne fait pas partie du programme, la formation biologique étant plus importante que la formation en informatique. Nous proposons que l'École Polytechnique ouvre ses portes aux projets de bioinformatique, et éventuellement que l'école ouvre un programme de Bioinformatique. Ainsi, que la formation des bioinformaticiens soit enrichie par une combinaison de cours en algorithmique et techniques d'optimisation, de statistiques et bio statistiques, et clairement de génie logiciel. La recherche en épidémiologie génétique serait donc dotée d'outils permettant d'orner la solution de problèmes complexes d'une construction systématique de logiciels permettant d'assurer la traçabilité des expériences et la fiabilité des solutions proposées.

CHAPITRE 10. RÉFÉRENCES

Abecasis, G. R., Cherny, S. S., Cookson, W. O., & Cardon, L. R. (2002). Merlin--rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet, 30*(1), 97-101.

Ailhaud, G., & Institut national de la santé et de la recherche médicale (2000). *Obésité : dépistage et prévention chez l'enfant*. Paris: INSERM.

Alberts, B. (2002). *Molecular biology of the cell* (4th ed.). New York: Garland Science.

Balding, D. J. (2006). A tutorial on statistical methods for population association studies. *Nature Reviews Genetics, 7*(10), 781-791.

Belmonte, M., & Yurgelun-Todd, D. (2001). Permutation testing made practical for functional magnetic resonance image analysis. *Ieee Transactions on Medical Imaging, 20*(3), 243-248.

Bielinski, S. J., Lynch, A. I., Miller, M. B., Weder, A., Cooper, R., Oberman, A., et al. (2005). Genome-wide linkage analysis for loci affecting pulse pressure: the Family Blood Pressure Program. *Hypertension, 46*(6), 1286-1293.

Bouchard, G., Roy, R., Casgrain, B., & Hubert, M. (1989). [Population files and database management: the BALSAC database and the INGRES/INGRID system]. *Hist Mes, 4*(1-2), 39-57.

Brunelle, P.-L. (2008). *Correction par simulation de tests multiples dans les études d'association génomique familiale.* Unpublished M.Sc.A thesis, Université de Montréal, Montréal.

Chakravarti, A. (2001). To a future of genetic medicine. *Nature, 409*(6822), 822-823.

Churchill, G. A., & Doerge, R. W. (1994). Empirical threshold values for quantitative trait mapping. *Genetics, 138*(3), 963-971.

Cowley, A. W., Jr. (2006). The genetic dissection of essential hypertension. *Nat Rev Genet, 7*(11), 829-840.

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., et al. (2007). mvtnorm: Multivariate Normal and t Distributions (Version 0.8-1): R package.

Gouyon, J.-B. (2005). Gènes et environnement: le partage des rôles. *Science & Vie, hors série, 230,* 58-63.

Griffiths, A. J. F., & Suzuki, D. T. (2002). *Introduction à l'analyse génétique / Anthony J. F. Griffiths ... [et al.] ; traduction de la 7e édition américaine par Chrystelle Sanlaville ; révision scientifique de Denise Aragnol et Dominique Charmot* (3e éd. ed.). Paris ; Bruxelles: DeBoeck.

Hamet, P., Merlo, E., Seda, O., Broeckel, U., Tremblay, J., Kaldunski, M., et al. (2005). Quantitative founder-effect analysis of French Canadian families identifies specific loci contributing to metabolic phenotypes of hypertension. *Am J Hum Genet, 76*(5), 815-832.

Haseman, J. K., & Elston, R. C. (1972). The investigation of linkage between a quantitative trait and a marker locus. *Behav Genet, 2*(1), 3-19.

Hines, W. W., Montgomery, D. C., Goldsman, D. M., Borror, C. M., Adjenge, L.-D., & Carmichael, J.-P. (2005). *Probabilités et statistique pour ingénieurs*. Montréal: Chenelière éducation.

Jones, E. O., T; Peterson, P (2001). SciPy : Open Source Scientific Tools for Python

Laird, N. M., Horvath, S., & Xu, X. (2000). Implementing a unified approach to family-based tests of association. *Genet Epidemiol, 19 Suppl 1*, S36-42.

Lange, C., van Steen, K., Andrew, T., Lyon, H., DeMeo, D. L., Raby, B., et al. (2004). A family-based association test for repeatedly measured quantitative traits adjusting for unknown environmental and/or polygenic effects. *Stat Appl Genet Mol Biol, 3*, Article17.

Moore, D., McCabe, G., Duckworth, W., & Sclove, S. (2003). The Practice of Business Statistics: Using Data for Decisions, Chapter 18: Bootstrap Methods and Permutation Tests. Available from http://bcs.whfreeman.com/pbs/cat_140/chap18.pdf.

Pichot, A. V., M. (2005). Hérédité: l'histoire d'un concept. *Science & Vie, hors série, 230,* 6-13.

Preiss, B. P. (2005). OPUS7. Data Structures and Algorithms with Object-Oriented Design Patterns in Python

R-Development-Core-Team (2007). R: A language and environment for statistical computing (Version 2.3.1). Vienna, Austria.

S.A.G.E. (2002). S.A.G.E. Statistical Analysis for Genetic Epidemiology, Release 42. (Version 4.2). Cleveland: Department of Epidemiology and Biostatistics, Rammelkamp Center for Education and Research, MetroHealth Campus, Case Western Reserve University: S.A.G.E.:.

Schalchli, L. (2005). Gène: un sens qui s'obscurcit. *Science & Vie, hors série, 230,* 30-39.

Sharma, S. (1996). *Applied multivariate techniques*. New York ; Toronto: J. Wiley.

Voet, D., & Voet, J. G. (1998). *Biochimie*. Paris: De Boeck Université.

Westfall, P. H., & Young, S. S. (1993). *Resampling-based multiple testing : examples and methods for P-value adjustment*. New York ; Toronto: Wiley.

Wigginton, J. E., Cutler, D. J., & Abecasis, G. R. (2005). A note on exact tests of Hardy-Weinberg equilibrium. *Am J Hum Genet, 76*(5), 887-893.

Yekutieli, D., & Benjamini, Y. (1999). Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *Journal of Statistical Planning and Inference, 82*, 171-196.

CHAPITRE 11. ANNEXES

ANNEXE 1. RÉSULTATS DES TESTS DE LIAISON DYNAMIQUE AVEC PERMUTATION DES GÉNOTYPES

Le tableau ci-dessous montre les résultats des tests de liaison dynamique d'un ensemble de SNPs sélectionnés selon sa signification dans une période spécifique, ainsi que la signification du test par période.

Tableau A.1 Liaison dynamique d'un ensemble de SNPs sélectionnés selon sa signification dans une période spécifique. L'étoile symbolise les SNPs étant localisés sur des régions chromosomiques proches à celles rapportées dans la littérature selon (Bielinski, et al., 2005).

SNP	*Trait*	*Valeur p empirique couché*	*Valeur p empirique debout*	*Chr*	*Position*	*Marshfield*	*Allèle A/B*	*Valeur p liaison dynamique mean_test*	*Valeur p liaison dynamique min_diff*	*Gène*
rs10494136	sv	4.40E-03	2.36E-01	1	111837755	145.7783215	C T	1.39E-02	6.04E-02	Ras-related protein Rap-1A
rs10494136	tpr	7.80E-03	3.31E-01	1	111837755	145.7783215	C T	1.33E-02	5.05E-02	Ras-related protein Rap-1A
rs10494170	tpr	1.19E-02	5.58E-01	1	115293759	148.2526527	A G	2.15E-02	7.91E-02	À proximité de TSPAN2 tetraspanin 2 et TSHB thyroid stimulating hormone
rs10489530	tpr	5.10E-03	5.77E-01	1	115551722	148.3972285	C T	2.12E-02	7.26E-02	À proximité de NGF nerve growth factor (beta polypeptide)
rs10494478	map	3.91E-02	4.00E-04	1	166073385	187.2563178	A G	9.88E-01	2.50E-02	BLZF1 basic leucine zipper nuclear factor 1

SNP	*Trait*	*Valeur p empirique couché*	*Valeur p empirique debout*	*Chr*	*Position*	*Marshfield*	*Allèle A/B*	*Valeur p liaison dynamique mean_test*	*Valeur p liaison dynamique min_diff*	*Gène*
rs10494478	sys	1.59E-01	5.50E-03	1	166073385	187.2563178	A G	9.35E-01	4.96E-02	BLZF1 basic leucine zipper nuclear factor 1
rs10489184	dia	1.75E-02	2.10E-03	1	166280698	187.4659525	A G	4.34E-01	5.58E-01	F5 coagulation factor V
rs10489184	map	4.00E-03	2.80E-03	1	166280698	187.4659525	A G	4.83E-01	8.29E-01	F5 coagulation factor V
rs10489184	sys	8.80E-03	1.30E-03	1	166280698	187.4659525	A G	6.00E-01	4.85E-01	F5 coagulation factor V
rs3753396	dia	9.90E-03	3.60E-03	1	193427399	*211.1537843	C T	6.02E-01	9.29E-01	complement factor H associé à la suceptibilité à l'infarctus du myocarde
rs3753396	map	7.00E-03	2.70E-03	1	193427399	*211.1537843	C T	6.95E-01	8.95E-01	complement factor H associé à la suceptibilité à l'infarctus du myocarde
rs3753396	sys	2.25E-02	7.00E-04	1	193427399	*211.1537843	C T	8.15E-01	2.05E-01	complement factor H associé à la suceptibilité à l'infarctus du myocarde
rs1934613	dia	4.00E-04	1.64E-02	1	207356392	*229.8608635	A G	7.01E-02	1.05E-02	KCNH1 potassium voltage-gated channel, subfamily H (eag-related), member 1
rs1934613	map	9.00E-04	6.70E-03	1	207356392	*229.8608635	A G	1.76E-01	6.03E-02	KCNH1 potassium voltage-gated channel, subfamily H (eag-related), member 1
rs1934613	sys	9.30E-03	4.31E-02	1	207356392	*229.8608635	A G	1.85E-01	1.03E-01	KCNH1 potassium voltage-gated

SNP	Trait	Valeur p empirique couché	Valeur p empirique debout	Chr	Position	Marshfield	Allèle A/B	Valeur p liaison dynamique mean_test	Valeur p liaison dynamique min_diff	Gène
										channel, subfamily H (eag-related), member 1
rs1934614	dia	1.19E-02	5.66E-02	1	207356607	*229.8610469	A T	4.52E-02	6.94E-02	KCNH1 potassium voltage-gated channel, subfamily H (eag-related), member 1
rs1934614	sv	9.98E-01	9.00E-04	1	207356607	*229.8610469	A T	8.75E-01	0.00E+00	KCNH1 potassium voltage-gated channel, subfamily H (eag-related), member 1
rs1934614	tpr	2.54E-01	1.80E-03	1	207356607	*229.8610469	A T	9.06E-01	1.96E-02	KCNH1 potassium voltage-gated channel, subfamily H (eag-related), member 1
rs206847	pr	6.00E-03	1.87E-02	2	31523018	49.36710239	A G	7.64E-02	2.93E-01	XDH xanthine dehydrogenase
rs395404	sv	2.22E-01	1.20E-03	2	40294277	60.60641069	C G	9.82E-01	7.57E-02	SLC8A1 solute carrier family 8 (sodium/calcium exchanger),
rs395404	tpr	4.52E-02	7.40E-03	2	40294277	60.60641069	C G	8.12E-01	2.61E-01	SLC8A1 solute carrier family 8 (sodium/calcium exchanger),
rs724333	dia	9.90E-03	4.16E-02	3	55933099	74.69949957	G T	2.37E-01	9.57E-02	ERC2 ELKS/RAB6-interacting/CAST family member 2
rs724333	map	9.00E-04	2.15E-01	3	55933099	74.69949957	G T	2.26E-01	5.00E-04	ERC2 ELKS/RAB6-interacting/CAST family member 2

SNP	*Trait*	*Valeur p empirique couché*	*Valeur p empirique debout*	*Chr*	*Position*	*Marshfield*	*Allèle A/B*	*Valeur p liaison dynamique mean_test*	*Valeur p liaison dynamique min_diff*	*Gène*
rs724333	sys	4.50E-03	4.26E-01	3	55933099	74.69949957	G T	2.24E-01	1.40E-03	ERC2 ELKS/RAB6-interacting/CAST family member 2
rs9311603	dia	3.70E-03	8.70E-03	3	56502406	75.58002467	A G	2.02E-01	1.69E-01	N.A
rs9311603	map	1.39E-02	7.22E-02	3	56502406	75.58002467	A G	2.60E-01	9.19E-02	N.A
rs3773714	dia	3.20E-03	2.42E-02	3	157742668	*171.4373105	A T	5.19E-01	5.04E-02	SSR3 signal sequence receptor, gamma (translocon-associated protein gamma)
rs10513494	dia	1.70E-03	5.00E-03	3	157773965	*171.4484626	A G	6.83E-01	1.77E-01	N.A
rs10513494	map	2.20E-03	7.60E-03	3	157773965	*171.4484626	A G	7.59E-01	1.58E-01	N.A
rs10513494	sys	1.15E-02	1.84E-02	3	157773965	*171.4484626	A G	7.84E-01	3.16E-01	N.A
rs10513688	dia	5.27E-02	1.40E-03	3	172209920	180.6009224	C T	8.63E-01	1.28E-01	SLC2A2 solute carrier family 2
rs10513688	map	6.76E-02	5.00E-03	3	172209920	180.6009224	C T	7.76E-01	2.05E-01	SLC2A2 solute carrier family 2
rs10513689	dia	4.26E-02	2.80E-03	3	172210360	180.6014207	C T	8.41E-01	2.58E-01	SLC2A2 solute carrier family
rs10513689	map	4.59E-02	4.90E-03	3	172210360	180.6014207	C T	7.34E-01	3.17E-01	SLC2A2 solute carrier family
rs7661217	tpr	2.70E-03	8.92E-01	4	47528992	60.39129483	C T	2.37E-01	5.00E-04	CORIN corin, serine peptidase
rs1453459	map	1.81E-02	2.13E-01	4	72828007	79.41821529	C T	8.74E-02	2.24E-02	N.A.
rs842873	sv	3.50E-03	2.21E-01	4	73061500	79.73198704	C G	3.30E-02	4.72E-02	N.A.
rs308388	map	1.62E-02	8.90E-01	4	124144684	125.7122185	C T	2.82E-01	3.00E-04	N.A.
rs1429119	map	1.78E-01	7.80E-03	4	148703144	146.910772	C T	9.74E-01	6.97E-02	N.A.
rs10519945	map	2.30E-01	8.10E-03	4	149478258	148.1246936	A G	9.76E-01	2.57E-02	NR3C2 nuclear receptor subfamily 3, group C, member 2

SNP	Trait	Valeur p empirique couché	Valeur p empirique debout	Chr	Position	Marshfield	Allèle A/B	Valeur p liaison dynamique mean_test	Valeur p liaison dynamique min_diff	Gène
rs10519945	sys	1.97E-01	3.30E-03	4	149478258	148.1246936	A G	9.84E-01	1.89E-02	NR3C2 nuclear receptor subfamily 3, group C, member 2
rs3846329	dia	3.90E-03	3.49E-02	4	149586943	148.4758447	G T	2.20E-01	4.76E-02	NR3C2 nuclear receptor subfamily 3, group C, member 2
rs10519959	dia	6.40E-03	1.50E-02	4	149627371	148.6064638	C T	2.23E-01	1.87E-01	NR3C2 nuclear receptor subfamily 3, group C, member 2
rs10519959	map	1.30E-03	6.00E-04	4	149627371	148.6064638	C T	5.00E-01	5.39E-01	NR3C2 nuclear receptor subfamily 3, group C, member 2
rs10519959	sys	4.03E-02	4.00E-04	4	149627371	148.6064638	C T	7.62E-01	1.21E-01	NR3C2 nuclear receptor subfamily 3, group C, member 2
rs4835136	dia	6.60E-03	1.53E-02	4	149627601	148.6072069	A G	2.38E-01	2.36E-01	NR3C2 nuclear receptor subfamily 3, group C, member 2
rs4835136	map	3.20E-03	1.30E-03	4	149627601	148.6072069	A G	6.14E-01	8.08E-01	NR3C2 nuclear receptor subfamily 3, group C, member 2
rs4835136	sys	5.05E-02	1.60E-03	4	149627601	148.6072069	A G	8.52E-01	1.24E-01	NR3C2 nuclear receptor subfamily 3, group C, member 2
rs2358469	map	2.38E-02	5.50E-03	4	149948186	149.6429871	C T	5.60E-01	7.84E-01	NR3C2 nuclear receptor subfamily 3, group C, member 2

SNP	Trait	Valeur p empirique couché	Valeur p empirique debout	Chr	Position	Marshfield	Allèle A/B	Valeur p liaison dynamique mean_test	Valeur p liaison dynamique min_diff	Gène
rs4077817	dia	8.40E-03	3.61E-02	5	95608129	104.6238381	C T	2.08E-01	8.49E-02	N.A
rs9295676	sys	1.56E-02	5.89E-01	6	26036355	43.25819418	G T	1.60E-02	1.27E-02	SLC17A2 solute carrier family 17 (sodium phosphate), member 2
rs3951042	dia	1.50E-03	4.18E-01	6	118829188	121.3563435	C T	4.56E-02	2.00E-04	N.A
rs3951042	map	4.00E-04	5.64E-01	6	118829188	121.3563435	C T	1.76E-02	1.00E-04	N.A
rs4027875	map	1.12E-02	6.57E-01	6	118829282	121.3563741	A G	1.03E-01	5.00E-04	N.A.
rs4027875	sys	7.20E-03	3.51E-01	6	118829282	121.3563741	A G	4.65E-02	4.10E-03	N.A.
rs10484287	map	6.40E-03	7.57E-01	6	118964872	121.4005784	A G	3.50E-02	0.00E+00	Proche à PLN phospholamban
rs7759088	map	5.10E-03	7.28E-01	6	119006062	121.4140069	A G	2.61E-02	2.00E-04	Proche à PLN phospholamban
rs7759088	sys	1.30E-02	6.05E-01	6	119006062	121.4140069	A G	1.51E-02	2.80E-03	Proche à PLN phospholamban
rs9320815	dia	2.00E-04	1.08E-02	6	121774189	122.3616983	C T	1.74E-01	5.50E-03	N.A
rs9320815	map	0.00E+00	4.45E-02	6	121774189	122.3616983	C T	8.74E-02	4.00E-04	N.A
rs9320815	sys	1.60E-03	1.30E-01	6	121774189	122.3616983	C T	7.57E-02	3.10E-03	N.A
rs16109	pr	7.00E-04	4.02E-01	7	24112609	38.66518316	A C	7.84E-02	1.20E-03	N.A
rs16089	pr	3.00E-04	8.07E-01	7	24121131	38.6683984	G T	1.55E-02	0.00E+00	N.A
rs6961069	tpr	5.72E-02	2.20E-03	7	79863612	93.24915793	A G	9.09E-01	1.99E-01	CD36 CD36 molecule (thrombospondin receptor)
rs7789369	tpr	1.84E-01	9.00E-04	7	79870798	93.25239365	A C	9.01E-01	4.61E-02	N.A.
rs10485996	map	1.17E-02	1.46E-02	7	94367436	107.0539459	A C	4.19E-01	3.69E-01	PPP1R9A protein phosphatase 1, regulatory (inhibitor)
rs10485996	sys	2.00E-03	2.69E-02	7	94367436	107.0539459	A C	2.45E-01	2.80E-02	PPP1R9A protein phosphatase 1,

SNP	Trait	Valeur p empirique couché	Valeur p empirique debout	Chr	Position	Marshfield	Allèle A/B	Valeur p liaison dynamique mean_test	Valeur p liaison dynamique min_diff	Gène
										regulatory (inhibitor)
rs854523	sys	4.10E-03	3.27E-02	7	94542884	107.2033403	A G	2.06E-01	5.50E-02	PPP1R9A protein phosphatase 1, regulatory (inhibitor)
rs1357319	dia	3.10E-03	4.90E-03	7	98954832	110.6765368	G T	6.07E-01	2.69E-01	CYP3A7 cytochrome P450, family 3, subfamily A, polypeptide 7
rs1357319	map	7.20E-03	7.90E-03	7	98954832	110.6765368	G T	7.91E-01	6.16E-01	CYP3A7 cytochrome P450, family 3, subfamily A, polypeptide 7
rs2037498	dia	5.30E-03	5.20E-03	7	98956201	110.6773825	C T	6.27E-01	3.26E-01	CYP3A7 cytochrome P450, family 3, subfamily A, polypeptide 7
rs2037498	map	9.70E-03	6.30E-03	7	98956201	110.6773825	C T	8.00E-01	7.08E-01	CYP3A7 cytochrome P450, family 3, subfamily A, polypeptide 7
rs2037499	dia	4.20E-03	5.00E-03	7	98957304	110.6780638	C G	6.27E-01	3.23E-01	CYP3A7 cytochrome P450, family 3, subfamily A, polypeptide 7
rs2037499	map	8.90E-03	7.50E-03	7	98957304	110.6780638	C G	7.97E-01	7.14E-01	CYP3A7 cytochrome P450, family 3, subfamily A, polypeptide 7
rs1732045	tpr	5.95E-01	1.00E-04	7	133619722	138.3320397	C T	9.48E-01	7.70E-03	N.A.
rs4733224	dia	7.50E-03	2.50E-02	8	31086911	58.58229799	A G	3.01E-02	1.60E-01	WRN Werner syndrome
rs349337	sv	8.00E-04	6.69E-01	8	73607217	92.47557556	A G	1.73E-02	9.30E-03	KCNB2 potassium voltage-gated channel

SNP	*Trait*	*Valeur p empirique couché*	*Valeur p empirique debout*	*Chr*	*Position*	*Marshfield*	*Allèle A/B*	*Valeur p liaison dynamique mean_test*	*Valeur p liaison dynamique min_diff*	*Gène*
rs349337	tpr	2.80E-03	9.28E-01	8	73607217	92.47557556	A G	1.70E-03	3.00E-03	KCNB2 potassium voltage-gated channel
rs2383869	sys	1.76E-02	2.01E-01	8	73729015	92.70340221	A G	1.65E-02	3.81E-02	KCNB2 potassium voltage-gated channel
rs2253667	tpr	5.60E-03	1.21E-01	8	73815202	92.87584358	A G	1.30E-01	1.07E-01	KCNB2 potassium voltage-gated channel
rs10505127	tpr	3.00E-04	1.00E+00	8	110122659	121.3011977	A C	2.94E-01	0.00E+00	N.A
rs3110037	tpr	1.00E-04	9.97E-01	8	110167364	121.3282511	C T	1.27E-01	0.00E+00	TRHR thyrotropin-releasing hormone receptor
rs10505126	tpr	5.10E-03	1.00E+00	8	110203456	121.3500924	G T	2.55E-01	0.00E+00	TRHR thyrotropin-releasing hormone receptor
rs303521	sv	3.60E-03	4.81E-01	10	90465841	109.239406	C T	6.00E-04	1.24E-02	N.A
rs7945727	map	7.26E-02	6.40E-03	11	100864321	98.56721649	A C	9.41E-01	1.75E-01	TRPC6 transient receptor potential cation channel
rs7945727	sys	7.68E-02	2.10E-03	11	100864321	98.56721649	A C	9.14E-01	8.23E-02	TRPC6 transient receptor potential cation channel
rs4129255	dia	1.28E-02	1.47E-02	11	100870413	98.57016714	C T	1.15E-01	4.31E-01	TRPC6 transient receptor potential cation channel
rs4129255	map	2.15E-02	1.80E-03	11	100870413	98.57016714	C T	5.95E-01	3.88E-01	TRPC6 transient receptor potential cation channel
rs10501981	map	5.85E-02	2.50E-03	11	100880825	98.57521016	C G	7.82E-01	1.52E-01	TRPC6 transient receptor potential cation channel
rs7925012	map	6.99E-02	5.70E-03	11	100888177	98.57877108	C T	8.40E-01	2.55E-01	TRPC6 transient receptor potential cation channel

SNP	*Trait*	*Valeur p empirique couché*	*Valeur p empirique debout*	*Chr*	*Position*	*Marshfield*	*Allèle A/B*	*Valeur p liaison dynamique mean_test*	*Valeur p liaison dynamique min_diff*	*Gène*
rs678815	map	3.30E-03	8.43E-02	11	102218987	99.22334543	C G	7.80E-03	1.44E-02	MMP3 matrix metallopeptidase 3
rs678815	sys	9.70E-03	1.15E-01	11	102218987	99.22334543	C G	2.68E-01	3.58E-02	MMP3 matrix metallopeptidase 3
rs1937388	tpr	2.20E-03	6.13E-01	13	77151203	60.40978622	C T	2.00E-04	1.04E-02	N.A
rs1924921	tpr	1.00E-03	7.26E-01	13	77360724	60.5818819	C T	1.20E-03	2.30E-03	Proche de EDNRB endothelin receptor type B
rs1946768	pr	7.60E-03	8.44E-01	16	70891871	88.31605212	C T	2.48E-02	2.10E-03	N.A.
rs17651134	pr	7.00E-03	3.80E-02	17	41406176	63.70202337	C T	2.70E-01	2.11E-01	MAPT microtubule-associated protein tau .
rs10514919	dia	1.44E-01	4.20E-03	17	42697128	*63.89068805	A C	8.68E-01	3.83E-02	The ITGB3 protein product is the integrin beta chain beta 3. Process associated: blood coagulation
rs10514919	map	1.02E-01	7.50E-03	17	42697128	*63.89068805	A C	8.18E-01	1.82E-01	The ITGB3 protein product is the integrin beta chain beta 3. Process associated: blood coagulation
rs10512510	sv	2.24E-01	2.00E-03	17	61798509	85.41514988	A T	9.57E-01	9.79E-02	PRKCA protein kinase C, alpha s
rs1799898	tpr	3.60E-03	3.38E-02	19	11088554	34.27308576	A G	1.80E-01	2.52E-01	LDLR low density lipoprotein receptor

ANNEXE 2. SPÉCIFICATION D'EXIGENCES LOGICIEL POUR LE TEST DE PERMUTATION POUR LA LIAISON DYNAMIQUE

ÉCOLE
POLYTECHNIQUE
MONTRÉAL

Prototype d'un logiciel pour l'analyse de liaison dynamique entre gènes candidats et phénotypes associes a la pression sanguine au cours de tests physiologiques dans les familles canadiennes françaises

Spécification d'exigences logicielles

Version 1.0
2007-07-11

Johanna Sandoval

Prototype d'un logiciel pour l'analyse de liaison dynamique entre gènes candidats et phénotypes associes a la pression sanguine au cours de tests physiologiques dans les familles canadiennes françaises	Version : 1.0
Spécification d'exigences logicielles	Date création: 2007-07-11

Historique des modifications du document

Date	Version	Description	Auteur
2007-07-11	1.0	Définition	Johanna Sandoval
2007-07-19	1.0	Envoie pour approbation et commentaires	Johanna Sandoval
2007-08-01	1.0	Envoie pour approbation et commentaires	Johanna Sandoval

Prototype d'un logiciel pour l'analyse de liaison dynamique entre gènes candidats et phénotypes associes a la pression sanguine au cours de tests physiologiques dans les familles canadiennes françaises	Version : 1.0
Spécification d'exigences logicielles	Date création: 2007-07-11

Table des matières

Prototype d'un logiciel pour l'analyse de liaison dynamique entre gènes candidats et phénotypes associes a la pression sanguine au cours de tests physiologiques dans les familles canadiennes françaises	Version : 1.0
Spécification d'exigences logicielles	Date création: 2007-07-11

Spécification d'exigences logicielles

1. Introduction

Ce document ci contient la définition du test de permutation appliqué à l'analyse de la composante dynamique de la liaison entre gènes candidats et phénotypes associés à la pression sanguine au cours de tests physiologiques. Il sert à clarifier la façon dont la méthode de permutation va être appliquée à la résolution du problème et peut contribuer à la détection préalable des défauts dans la définition ou la compréhension du problème ainsi qu'à établir une base pour la définition des planes de validation et vérification.

1.1 Objectif du document

Cette spécification décrit le comportement attendu de l'application et les facteurs nécessaires à son implantation. Il est souhaitable que les recommandations sur la façon d'approcher le problème et la méthodologie utilisée, ainsi que les outils sélectionnés et les contraintes définies soient discutées avant de commencer l'implantation.

1.2 Portée du document

Cette définition-là n'explique pas le problème en soi mais l'application qui sera construite pour le résoudre. L'application est modelé en fonction des fonctionnalités que la composent.

Pour continuer, le but de cette étude est de révéler la dynamique de la liaison génétique entre un ensemble de SNPs donné associés à des gènes candidats et certains phénotypes spécifiques qui entre en jeu dans la détermination de la pression artérielle (PA) un fois que les individus sont soumis à des tests physiologiques. D'autres méthodes ont été implantées pour expliquer les cas où certains SNPs sont liés aux mêmes phénotypes seulement durant certaines périodes. Enfin, la comparaison des résultats des tests antérieurs et ceux produits pour cette application ne font pas partie de la portée du projet.

1.3 Définitions, acronymes et abréviations

SNP: Single Nucleotide Polymorphism.
PA : Pression Arterielle

1.4 Références

Belmonte,M; Yurgelun-Todd,D. Permutation Testing Made Practical for Functional Magnetic Resonance Image Analysis. *IEEE Transactions on Medical Imaging* **20**(3):243-248 (March 2001).

Garson, G. David (n.d.). "Correlation". *Statnotes: Topics in Multivariate Analysis.*1998-2005. [En ligne]. Disponible: http://www2.chass.ncsu.edu/garson/pa765/statnote.htm. [Consulté le 11 juin 2007].

Desmarais, M. « LOG 4315, Atelier de génie logiciel ». École Polytechnique de Montréal Département de

Génie informatique, 2005. [En ligne]. Disponible: http://www.cours.polymtl.ca/log4315/gabarits.php. [Consulté le 11 juillet 2007].

IEEE Std 830-1998, IEEE Recommended Practice for Software Requirements Specifications. 1998.

1.5 Vue d'ensemble

Le document qui suit contient la caractérisation et les contraintes pour l'implantation du prototype en question. Cette spécification a été construite selon le standard IEEE STD 830-1998. La description général de la section 2 expose comment les interfaces utilisateur, système et logicielle ont été désignées, les contraintes reliées à l'utilisation de mémoire, les fonctions, et les conditions générales pour que l'implantation soit fonctionnelle. La section 3 contient un modèle décrivant les paramètres, les interfaces (fichiers d'entrées et sorties), les fonctionnalités et les contraintes impliqués. Les contraintes de conception et la sécurité se résument dans les sections 4 et 5. Les exigences de documentation sont notés dans la section 6. Il est a souligner que aucune norme n'est applicable pour la construction et les tests du prototype. C'est pourquoi la section de normes applicables a été exclue. La section 7 contient une classification des exigences fonctionnelles préalablement définies.

2. Description générale

L'application prend un fichier contenant les valeurs t résultantes du test de liaison des phénotypes associés à la PA et ses phénotypes intermédiaires et un fichier contenant la série de temps idéale. Ensuite, les corrélations entre ces deux séries de temps par phénotype et par SNP (chaque phénotype constitue un expérience différente) sont calculées. Selon l'assomption d'absence de corrélation entre la série de temps et la série idéale, une distribution de la corrélation maximale obtenue est construite en répartissant aléatoirement la série de temps des valeurs t. Les corrélations des séries originales sont ajustées par rapport à la distribution nulle et la valeur de probabilité calculée. Cette dernière explique la relation entre la série de temps des tests de liaison et de physiologie.

2.1 Perspectives du produit

2.1.1 Interfaces système

Le prototype est une application indépendant (stand-alone), ne requiert pas de communication avec d'autres systèmes. L'appel à d'autres modules en python se fait directement. L'interface avec le logiciel R est pourvue par la librairie rpy de Python.

2.1.2 Interfaces utilisateurs

Aucune interface graphique n'est planifiée. L'utilisateur du logiciel doit connaître la localisation des fichiers contenant les valeurs t, la série de temps idéale et le nom et la localisation fichiers de sortie et de log.

2.1.3 Interfaces matérielles

Aucune interface matérielle n'est planifiée.

2.1.4 Interfaces logicielles

Le langage de programmation doit posséder une interface avec l'application R, qui serait utilisé pour les calculs de la corrélation et d'autres calcules statistiques pertinentes. Des composants développés par Pierre-Luc Brunelle pour des tâches telles que la création de dossiers, l'accès aux données sur la base de données CARDIO6, l'utilisation d'expressions réguliers et d'autres fonctions miscellanées pourraient être éventuellement utilisés.

2.1.5 Interfaces de communication

Des interfaces de communication ne sont pas prévues.

2.1.6 Contraintes de mémoire

Les structures de données nécessaires sont classifiées selon qu'elles appartiennent à la mémoire disque (secondaire) ou primaire. Voici leur description :

Mémoire disque (secondaire):

- Fichier texte contenant la liste des valeurs t, dont la structure doit contenir : phénotypes, SNPs, time point, valeur t et valeur p (taille = 4.5 Mb);
- Fichier texte contenant les résultats du test contenant : la liste de SNPs définis comme « dynamiquement liés », dont la structure contient: phénotypes, SNPs, valeur de la corrélation originale, valeur de la statistique et valeur p (taille estimée = 4 Mb);
- Fichier texte contenant la série de temps idéale qui représente les tests physiologiques (un signal carré contenant des valeurs différentes pour chaque condition, c'est-à-dire couché, debout, test mathématique, etc… (taille estimée : 1Kb);
- LOG des résultats calculés représente une partie importante des besoins en espace sur le disque. Les messages doivent être contrôlés pour ne pas excéder 1 GB (disponibilité de 3Gb sur les serveurs de calculs), soit 100 MB par itération.

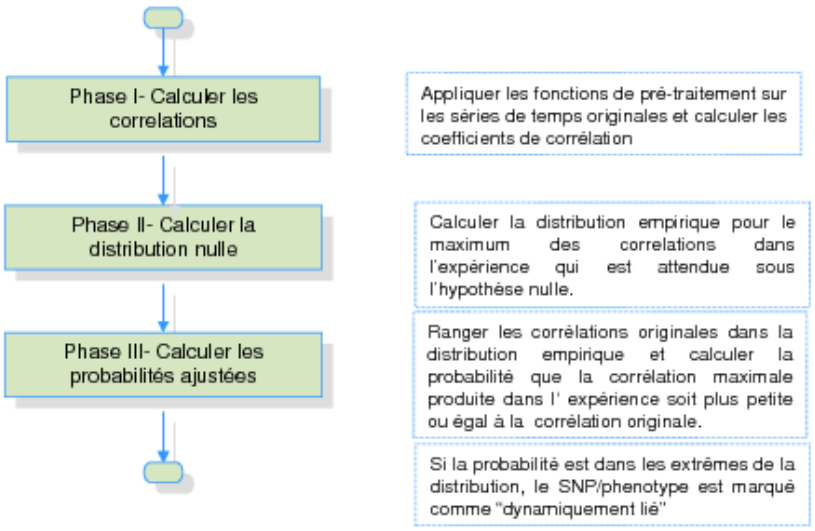
Mémoire primaire :

- Liste des SNPs dans le fichier des valeurs t;
- Liste des phénotypes dans le fichier des valeurs t;
- Dictionnaires contenant les séries de temps (valeurs t et série idéale). La série de temps doit être accessible par SNP et temps;
- Arbre binaire contenant les corrélations originales ordonnées(clé par la valeur de la corrélation et par le SNP qui l'a produit);
- Arbre binaire avec les corrélations maximales (clé par la valeur de la corrélation et par le SNP qui l'a produit);

- Arbre binaire contenant les corrélations substituts ordonnés (clé par la valeur de la corrélation et par le SNP qui l'a produit). Les auteurs recommandent au maximum trois corrélations par itération.

2.2 Fonctions du produit

On peut distinguer trois fonctions, qui équivalent à chaque phase de l'algorithme :



2.3 Caractéristiques des utilisateurs

L'utilisateur n'a pas besoin d'une immense expertise technique. En fait, l'interaction de l'application avec les usagers est contrainte à la définition des paramètres de l'application.

2.4 Contraintes

L'application sera disponible sur LINUX.

2.5 Hypothèses et dépendances

R doit être préalablement installé sur la machine utilisée pour le calcul. De plus, les librairies nécessaires seront celles installées par défaut (base, stats, etc).
Similairement, Python doit être préalablement installé sur le serveur de calcul. L'interface rpy doit être disponible en Python.

2.6 Exigences reportées

D'autres fonctions de dépuration de la série de temps telle que l'élimination de l'auto corrélation est recommandé cependant, étant donné que notre série de temps est courte, l'application des modèles d'auto-régression ou des moyennes mobiles s'avère inutile.

3. Exigences spécifiques

3.1 Interfaces externes

3.1.1 Paramètres de l'application

Paramètres	Validations
Le nom et la localisation du fichier contenant les valeurs t	Le fichier doit exister; pour les doublets SNP/Phénotype dont la valeur t n'existe pas, la valeur doit être considérée comme nulle (valeur manquante).
Le nom et la localisation du fichier contenant la série de temps idéale	Le fichier doit exister; une valeur représentant le test physiologique doit exister pour chaque temps enregistré dans le fichier des valeurs t.
Le nom et la localisation du fichier de sortie contenant les corrélations, les valeurs p et l'indicateur de liaison dynamique par SNP et Phénotype	Si le fichier existe il doit être re-écrit.
Le seuil utilisé pour les tests de corrélation	Une chiffre décimal dont la valeur varie entre 0 et 1.
Le nombre d'itérations pendant la phase II, ce qui équivaut à la taille de la distribution empirique	La valeur recommandée est 10000.

3.1.2 Fichiers d'entrée

Valeurs t: Fichier en format texte séparé par tabulations contenant SNP, Phénotype, temps de la mesure, valeur t, valeur p.
Série de temps idéale: Fichier en format texte séparé par tabulations contenant tous les temps consignés au fichier des valeurs t et la valeur associée au test physiologique (soit 0 pendant la position couchée, 1 pour la position débout ou 2 pour le test de stress mental).

3.1.3 Fichiers de sortie

Résultats : Fichier en format texte séparé par tabulations contenant le SNP, le phénotype, la valeur de la corrélation originale et ajustée, la probabilité de la corrélation et l'indicateur de liaison dynamique.

Fichiers de résultats intermédiaires des phases I et II, qui enregistrent les données de chaque opération, permettent à l'application d'initier à partir d'une étape intermédiaire.

LOG : Sommaire des opérations indiquant les opérations importantes de l'application :

- Le SNP choisi et la corrélation maximale dans la phase II;
- Les corrélations substitues;
- Les SNPs qui sont enlevés pendant la phase III et l'itération durant laquelle ces SNPs sont éliminés;
- Messages d'erreur des paramètres et inclusion de données manquantes;
- Le nombre d'itération et les valeurs de la série de temps permutées.

3.2 Fonctionnalités

Les trois fonctions suivantes équivalent à chaque phase de l'algorithme. Il est à souligner que les phases I et II peuvent être exécutés parallèlement contrairement à la phase III qui doit nécessairement les suivre.

3.2.1 Phase I- Calculer les correlations sur les séries originales

Durant la première phase, on applique des filtres afin d'éliminer la tendance des séries originales. L'élimination de la tendance consiste à calculer la courbe de régression linéaire des séries de temps par rapport à ceux de mesure et à soustraire cette courbe de la série de temps original (En R: resid(lm(y ~ times))). D'autres sortes d'élimination de la tendance sont envisageables et pourraient être appliquées dans le but de comparer les résultats. Par exemple, enlever la moyenne de la série de temps.

L'étape suivante consiste à calculer la corrélation entre tous les séries de temps (par SNP) et la série de temps idéale. Le coefficient de corrélation choisi est le *Tau* de Kendall, communément utilisé pour corréler deux variables ordinales ou une variable continue avec une ordinale (ce qui est notre cas). La valeur p du coefficient de corrélation doit être continuellement vérifiée. Si la corrélation n'est pas significative (α=0.05), la valeur 0 sera mise dans la liste pour éviter que la corrélation pour le SNP fasse partie de la distribution maximale.

3.2.2 Phase II- Calculer la distribution nulle

La distribution nulle de la corrélation maximale attendue sous l'hypothèse nulle pour l'ensemble des SNPs est calculée en répétant les étapes suivantes 10000 fois :

- La séquence de temps est répartie au hasard pour tous les SNPs.
- En utilisant cette séquence, les corrélations avec la série de temps idéale sont calculées.
- La corrélation ayant la plus grande la valeur absolue (la magnitude) est enregistrée dans une liste ordonnée avec le SNP qui l'a produite. Selon le paramètre indiquant le nombre de SNPs additionnels que l'on récupère, les corrélations suivant la maximale et les SNPs qui l'ont produits sont ainsi notés dans une liste séparée (substitues).

3.2.3 Phase III- Calculer les probabilités ajustées

D'abord, la liste de probabilités pour chaque SNP est initialisée avec la valeur 0.5 pour tous les SNPs.

Ensuite, la liste de correlations produite dans la Phase I est ordonnée. On extrait itérativement la corrélation maximale avec le SNP qui l'a produite. La corrélation est classée dans la distribution empirique calculée dans la phase II. Ce classement est projeté sur l'intervalle [0,1] afin de trouver la probabilité que la corrélation maximale produite, dans un espace de SNPs non liés dynamiquement, soit plus petite ou égale à celle produite pour ce SNP-ci. Si la probabilité ajustée est plus petite ou égale que le seuil α (par défaut 0.05, mais il doit être paramétrable) divisé par deux ou 1-α/2, alors le SNP est marqué comme activé. D'un autre côté, les valeurs de la corrélation maximale pour ce dernier sont effacées de la distribution nulle construite dans la phase II et sont remplacées pour les corrélations substitues.

Il est à noter que ce procédé-ci est répété jusqu'à ce que la probabilité ajustée de la corrélation maximale extraite devient non significative.

Enfin, un fichier de sortie doit être créé contenant la liste de SNPs, la valeur de la corrélation originale, la valeur de la corrélation ajustée, la probabilité calculée dans la phase III et l'indicateur de liaison dynamique (oui ou non).

3.3 Exigences supplémentaires

3.3.1 Utilisabilité

Aucune exigence d'utilisabilité n'est requise.

3.3.2 Fiabilité

Voici la liste des anomalies qui seront contrôlées au cours de cette analyse:

Anomalies	Criticité	Procédures à suivre
Erreur dans le calcul de la corrélation de la phase I.	Mineur	Control par exception, permettre au programme de continuer
Erreur dans le calcul de la corrélation de la phase II.	Mineur	Control par exception, permettre au programme de continuer
Fautes indétectables durant une itération de la Phase II	Mineur	Control par exception, permettre au programme de continuer

3.3.3 Performance - Mode d'opération lors de la dégradation

L'application doit être capable de recommencer d'une itération donnée. Alors, il faut qu'elle soit capable de reprendre à partir des fichiers de résultats intermédiaires des phases I et II. Ainsi, si la performance du système se détériore, l'utilisateur pourra terminer abruptement l'application (le programme débutera au dernier point

d'exécution des phases I ou II).

3.3.4 Maintenabilité

Aucune exigence de maintenabilité

4. Contraintes de conception

4.1 Langage de programmation

Le langage de programmation choisi est Python. Par la suite, l'utilisation des structures de donnés spécifiques au problème sera exécuté en C++. Ultimo, le calcul des corrélations sera opéré en R.

5. Sécurité

Aucune mesure de sécurité est prévue; les mesures de protection de l'information traitée sont les mêmes que celles qui existent pour l'accès aux données.

6. Exigences de documentation d'utilisateur et d'aide en ligne

La définition du problème, les résultats et son interprétation seront consignés dans un mémoire de maîtrise. La documentation de l'utilisateur ne contiendra que des instructions pour l'exécution de l'application.

7. Classification des exigences fonctionnelles

Fonctionnalité	Type
Phase I	Essentielle
Phase II	Essentielle
Phase III	Essentielle
Mode d'opération lors de la dégradation	Souhaitable

Fonctionnalité	Type
Phase I	Essentielle
Phase II	Essentielle
Phase III	Essentielle
Mode d'opération lors de la dégradation	Souhaitable